Formation

Plan de Gestion des Données (PGD)

Sandrine Auzoux, Pauline Corbière, Laurence Dedieu

8/9 avril et 11/12 avril 2019

Programme de la formation

Jour 1

Accueil, présentation des participants

1. Définitions en matière de données, PGD et cadre juridique général
2. Enjeux du PGD
3. Rédiger le PGD
4. Décrire ses données

   Application sur vos données

Jour 2

5. Sécuriser, stocker et archiver ses données

   Application sur vos données

6. Bonnes pratiques juridiques

   Application sur vos données

7. Partage et valorisation des données

   Application sur vos données

Quizz et évaluation

1

DÉFINITIONS

1. Définitions clés
→ Donnée - Information - Connaissance

Donnée

Résultat direct d'une mesure (faits, observations, éléments bruts)

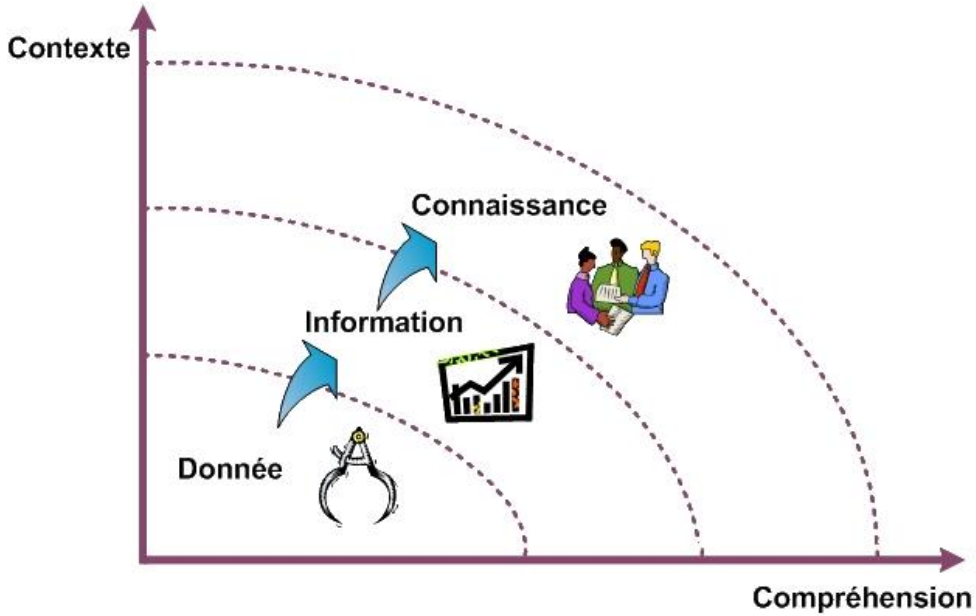
Ex. : 40°C

Information

Donnée interprétée

(qui, quoi, quand, où)

Ex. : température de l'air en degré Celsius à 14h à Montpellier



Connaissance

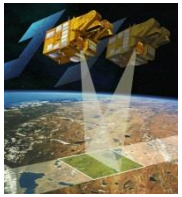
Information comprise (comment, pourquoi)

Ex. : il fait chaud. C'est l'été.

1. Définitions clés
→ Les données de la recherche

Données d'observation

Données capturées en temps réel généralement uniques et impossibles à reproduire. Elles ont vocation à être conservées de façon pérenne.

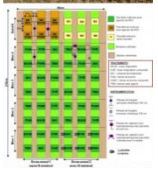


télédétection

enquêtes

Données expérimentales

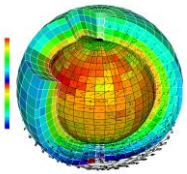
Données obtenues à partir d'équipements en laboratoire, suivant une méthodologie bien définie. Potentiellement reproductibles, mais à des coûts parfois prohibitifs.



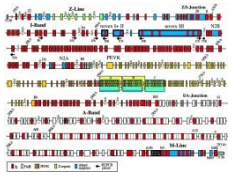
Résultats agronomiques

Données de simulation

Données générées à partir de modèles.



Modèle climatique



Séquence de gènes

Données dérivées ou compilées

Données résultant du traitement ou de la combinaison de données «brutes».

Data mining – text mining

Données de référence

Collection de jeux de données qui ont été revus, annotés et mis à disposition par les pairs.

1. Définitions clés
→ Les métadonnées

Les métadonnées sont des « *données qui décrivent des données* » :

- **Information** structurée associée à un "objet", un document ou un jeu de données
- **Documentation** qui permet à l'utilisateur de comprendre, de comparer et d'échanger le contenu du jeu de données décrit

Un objet sans étiquette n'est connu que de son auteur

Il existe des **standards** de métadonnées :

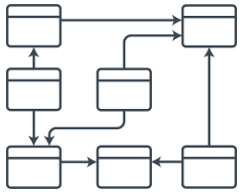
- Standards minimaux (ex : Dublin Core)
- Standards métiers (ex : EML, DDI…)

Il est conseillé de produire les métadonnées au **moment de la collecte ou de la création** des données plutôt qu'à posteriori. Les métadonnées seront **complétées tout au long du cycle de vie** des données.

1. Définitions clés
→ Les bases de données

Une base de données est un **ensemble de données** stockées de façon :

- **Persistante** : stockage permanent
- De **redondance minimale** : la même information n'est idéalement présente qu'une fois (unicité)
- **Exhaustive** : la base de données contient toutes les informations requises pour le service attendu
- **Structurée** : la structure est définie dans un schéma (le « modèle »)



- Elle est gérée par un système de gestion de bases de données

ORACLE

neo4j

PostgreSQL

mongoDB

MySQL

Microsoft SQL Server

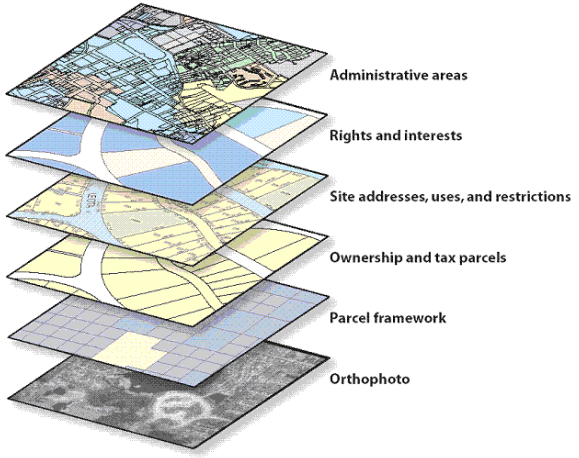
Microsoft Access

1. Définitions clés
→ Les jeux de données

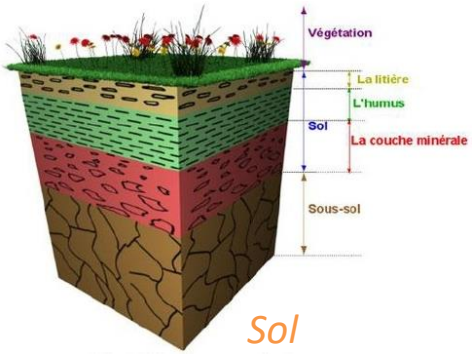
Un jeu de données (*dataset*) est l'agrégation d'enregistrements de données organisés pour former un **ensemble cohérent**. Les jeux de données numériques sont formatés de telle sorte qu'ils soient communicables, interprétables et adaptés à un traitement informatisé.

Météo

Cartographie d'acteurs

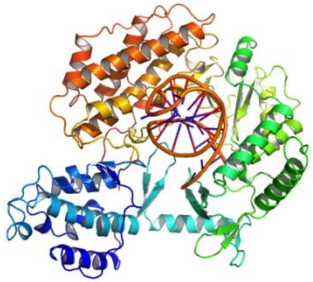


Couches SIG



Sol

Culture



Biologie cellulaire

Séries temporelles

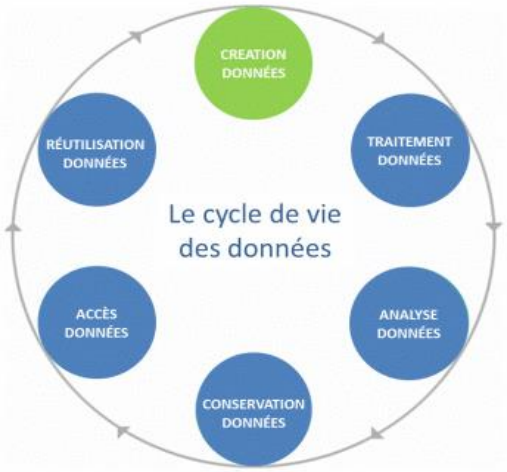
1. Définitions clés
→ Le plan de gestion des données

Le PGD est un document qui :

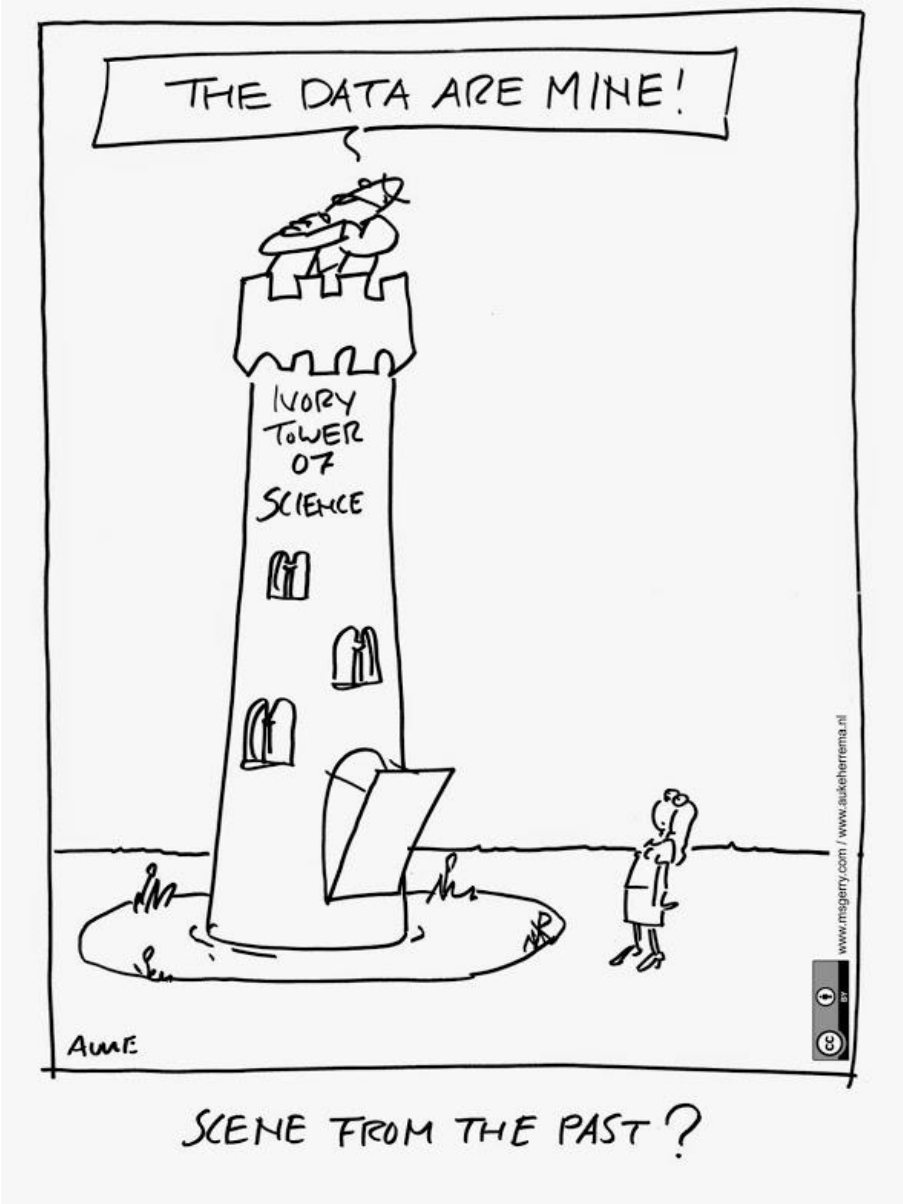
- Rassemble les **règles de gestion** et de **documentation** des données produites et réutilisées au cours d'un projet de recherche
- Renseigne les modalités de **partage**, de **conservation** et de **valorisation** des données après la clôture du projet
- Favorise la **compréhension**, la **diffusion** et la **réutilisation** des données

Il doit être **mis à jour** tout au long du projet : 3 versions minimums en début, milieu et fin de projet.

Il s'appuie sur le **cycle de vie des données**.



1. Les données vues par les juristes



1. Définitions juridiques
→ Les données

- *Il n'existe pas de définition légale de la donnée*

- Arrêté du 22 décembre 1981 *sur l'enrichissement de la langue française : représentation d'une information conventionnelle destinée à faciliter son traitement*

- Définition de l'OCDE des données de la recherche :
  Les données de recherche sont définies comme des enregistrements factuels (chiffres, textes, images et sons) qui sont utilisés comme sources principales pour la recherche scientifique et sont généralement reconnus par la communauté scientifique comme nécessaires pour valider des résultats de recherche. Un ensemble de données de recherche constitue une représentation systématique et partielle du sujet faisant l'objet de la recherche.

- *D'un point de vue juridique il n'y a pas de différence entre données brutes, élaborées ou métadonnées.*

1. Définitions clés du point de vue juridique

La donnée et le droit

- Par principe les données ne sont pas appropriables par un droit de propriété matérielle ou immatérielle
- La donnée est de libre parcours
- La donnée en elle-même ne fait pas l'objet de protection

⚠ **Exceptions**

La donnée peut faire l'objet d'une protection par un droit de propriété intellectuelle

Exemple : protection par le droit d'auteur (dessin, photographie…), droit des marques

Limitation de l'usage/Nécessité d'obtenir une autorisation expresse (licence)

⚠ Les données contenues dans **une base de données**

La base de données

Définition : La base de données est définie à l'article L112-3 du Code de la Propriété Intellectuelle :

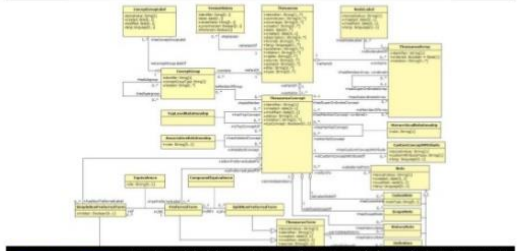
*« On entend par base de données un **recueil** d'œuvres, **de données** ou d'autres éléments indépendants, **disposés de manière systématique** ou **méthodique**, et **individuellement accessibles** par des **moyens électroniques** ou par **tout autre moyen**. »*

Quelques exemples de bases de données – un fichier produit avec un tableur

Quelques exemples de bases de données – un fichier papier

Cour d'Appel de Paris, 12 septembre 2001 : Peu importe que la base de données soit sous une forme papier, le support est indifférent.

Quelques exemples de bases de données – un thesaurus



Arrêt Jeuxvideo.com : « *Une liste alphabétique de mots n'est pas protégée par le droit d'auteur, si elle constitue une compilation sans originalité. En revanche, une liste de mots présentant une structure originale peut être protégée par le droit d'auteur. Un thésaurus par exemple, c'est-à-dire une liste hiérarchisée de termes soigneusement sélectionnés est protégeable* ».

1. Définitions clés du point de vue juridique

Cadre juridique de la base de données

La base de données peut être appréhendée sous 3 angles :

A. **Les données** contenues dans la base de données (le contenu)
B. **La structure** de la base de données (le contenant)

C. La base de données dans son **ensemble** (contenant + contenu)

A) La donnée

Le principe

Les données sont de libre parcours et ne sont pas appropriables.

Les exceptions

1. Certaines données peuvent bénéficier de la protection par le **droit d'auteur** (article L111-1 et s. du code de la propriété intellectuelle) Il s'agit de données constituant des œuvres de l'esprit : dessins, vidéos, plans, photographies, croquis, etc….

   ❖ **Condition** : l'œuvre doit être originale

   - Notion d'originalité : selon la CJCE, il y a création intellectuelle propre à son auteur lorsque l'auteur a pi exprimer son esprit créateur de manière originale.

   Approche classique : l'originalité est le reflet de la personnalité de l'auteur.

   ❖ Conséquences : pour utiliser ces données il faut une autorisation expresse écrite et préalable de l'auteur de la donnée protégée (exception : la courte citation).

   En pratique : à l'exception des photos et dessins, les données bénéficiant d'une protection par le droit d'auteur sont assez rares dans le monde de la recherche.

1. Définitions clés du point de vue juridique

B) LA STRUCTURE

La structure ou architecture de la base de données peut faire l'objet d'une **protection par le droit d'auteur**.

- Condition : la structure doit être **originale**
- Pas évident de retrouver l'empreinte de la personnalité de l'auteur dans une base de données
- Critère du législateur : art.L122-3 : la structure est protégeable lorsque « par le choix ou la disposition des matières, elles constituent des créations immatérielles »
- Critère retenu par les juges : empreinte de la personnalité de l'auteur est caractérisée par les choix opérés pour mettre au point la structure ; parfois le critère de la nouveauté est retenu

En pratique : posez vous la question de la paternité d'une structure lorsque vous utilisez une structure pré-existante.

1. Définitions clés du point de vue juridique

C) La base de données dans son ensemble

- Le droit du producteur de la base de données est le **droit d'interdire l'extraction et/ou la réutilisation de tout ou partie du contenu de la base**

- Objectif : Lutter efficacement contre le pillage des bases de données, à la portée de tous notamment avec le développement du numérique

- Protéger l'initiative et l'investissement réalisé pour mettre au point une base de données

- Transposition de la directive du 11 mars 1996 dans la loi 1er juillet 1998

- Articles L.341-1 et suivants du code de la propriété intellectuelle

1. Définitions clés du point de vue juridique

La base de données

Date	Lieu	Taille	Photo	Climat	Température moyenne
2015	Brésil	240 cm		Humide	26°
2014	Afrique du Sud	610 cm		Sec	32°
2013	Togo	460 cm		Sec à très sec	29°
2012	Portugal	420 cm		Humide à sec	28°
2011	Côte d'Ivoire	800 cm		Sec	24°

1. Définitions clés du point de vue juridique

La structure

Date	Lieu	Taille	Photo	Climat	Température moyenne
2015					
2014					
2013					
2012					
2011					

1. Définitions clés du point de vue juridique

Les données

- 2015, 2014, 2013, 2012, 2011
- Brésil, Afrique, Togo, Portugal

Pas de protection Données de libre parcours

- 240 cm, 610 cm, 460cm, 420 cm, 800 cm
- Photos





Exception : Droit d'auteur sur les photographies

1. Définitions clés du point de vue juridique

Titularité des droits

Droit du producteur de bases de données

Celui qui prend l'initiative et le risque financier

En pratique : la personne morale, l'employeur, le bailleur, le partenaire

- Il peut y avoir plusieurs co-producteurs

Droit d'auteur sur la structure

Personne physique qui « écrit » la structure

En pratique : le salarié chercheur, informaticien, chef de projet

- Il peut y avoir plusieurs auteurs

Données

Pas de droit = pas de titularité sauf si la donnée bénéficie de la protection par le droit d'auteur = personne physique ayant réalisé l'œuvre

Ex : chercheur ayant réalisé la photographie

FOCUS

Données = Résultats

La définition des « Résultats » dans les contrats de :

- ✓ Collaboration de recherche :
- ✓ Partenariat ;
- ✓ Consortium ;
- ✓ Prestation de service…

Le sort des données, de l'utilisation qui peut en être faite est souvent déterminée contractuellement = lire les clauses relatives aux résultats issus des projets dans les contrats

Article 1103 du code civil : les contrat légalement formés tiennent lieu de loi à ceux qui les ont faits.

2

LES ENJEUX DU PGD

2. Les enjeux du PGD

La réutilisation des données

Constat alarmant de gaspillage de la recherche

90% des données non réutilisables car stockées sur disques durs locaux

70% des chercheurs disent avoir essayé de reproduire des expérimentations sans succès (Nature 2016) https://www.nature.com/polopoly_fs/1.19970!/menu/main/topColumns/topLeftColumn/pdf/533452a.pdf

50% littérature biomédicale fausse (Lancet 2015) https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(15)60696-1/fulltext?code=lancet-site

89% des études animales citées pour les essais chez l'homme non publiées (PLoS Biol 2018; Science 2018) http://www.h2mw.eu/redactionmedicale/2018/04/cet-article-de-plos-biology-intitul%C3%A9-preclinical-efficacy-studies-in-investigator-brochures-do-they-enable-riskbenefi.html

Les enjeux

- Minimiser les risques de perte de données
- Assurer la reproductibilité de la recherche
- Éviter les duplications d'expériences
- Optimiser le financement de la recherche (meilleur usage de l'argent public)

2. Les enjeux du PGD

La réutilisation des données

- Assurer transparence, intégrité, traçabilité, fiabilité
  - → meilleure garantie contre la fraude scientifique
  - → renforce la confiance des citoyens
- Données = bien public
- Faire avancer la science et accélérer l'innovation

➢ vos données peuvent avoir un potentiel de réutilisation dans vos domaines ou dans d'autres et servir à:

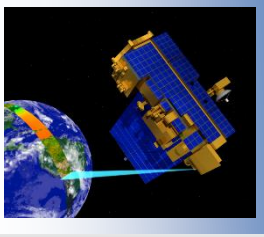
- Paramétrer des modèles
- Créer de nouvelles applications ou services
- Réaliser des méta-analyses → changements d'échelle spatiale, temporelle
- Enrichir les données produites par d'autres et contribuer aux travaux globaux sur changement climatique, santé publique, protection ressources naturelles, …

Modèle de prédiction des voies migratoires d'oiseaux développé à partir de diverses données accessibles

eBird

Land Cover

Meteorology



MODIS – Remote sensing data

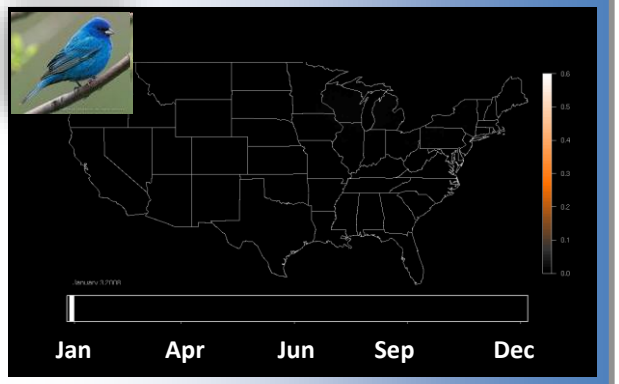
TeraGrid

$$F(X,s,t) = \frac{1}{n(s,t)} \sum_{i=1}^{m} f_i(X,s,t) I(s,t \in \theta_i)$$

Spatio-Temporal Exploratory Models predict the probability of occurrence of bird species across the United States at a 35 km x 35 km grid.

Model results

Occurrence of Indigo Bunting (2008)



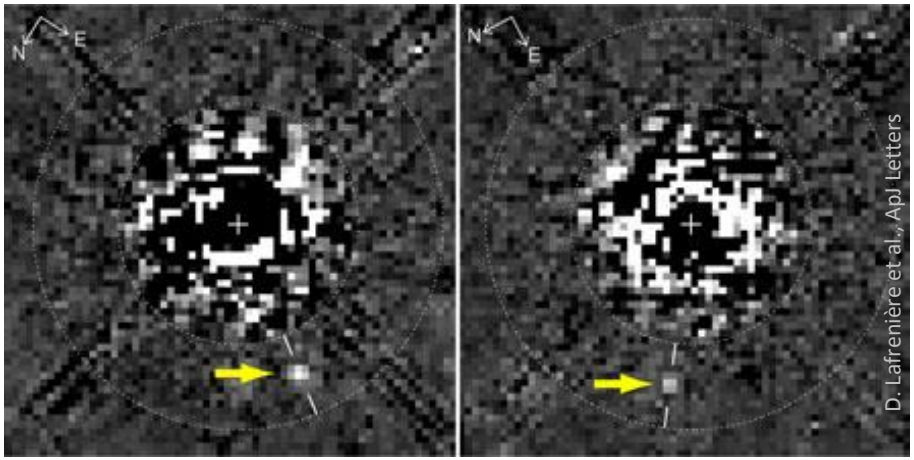
Potential Uses-

- Examine patterns of migration
- Infer impacts of climate change
- Measure patterns of habitat usage
- Measure population trends

DataONE Education Module: Data Management. DataONE. Retrieved August 22, 2018. From http://www.dataone.org/sites/all/documents/L01_DataManagement.pptx

Potentiel des données de recherche pour le future
grace à l'évolution des techniques et méthodes d'analyse

A new image processing technique reveals something not before seen in this Hubble Space Telescope image taken 11 years ago: A faint planet (arrows).



"Planet hidden in Hubble archives"
Science News
(Feb. 27, 2009)

It tells <u>how valuable maintaining long-term archives can be</u>

"Your well-managed and accessible data can contribute to science in ways you may not even imagine today! "

DataONE Education Module: Data Management. DataONE. Retrieved August 22, 2018.
From http://www.dataone.org/sites/all/documents/L01_DataManagement.pptx

2. Les enjeux du PGD

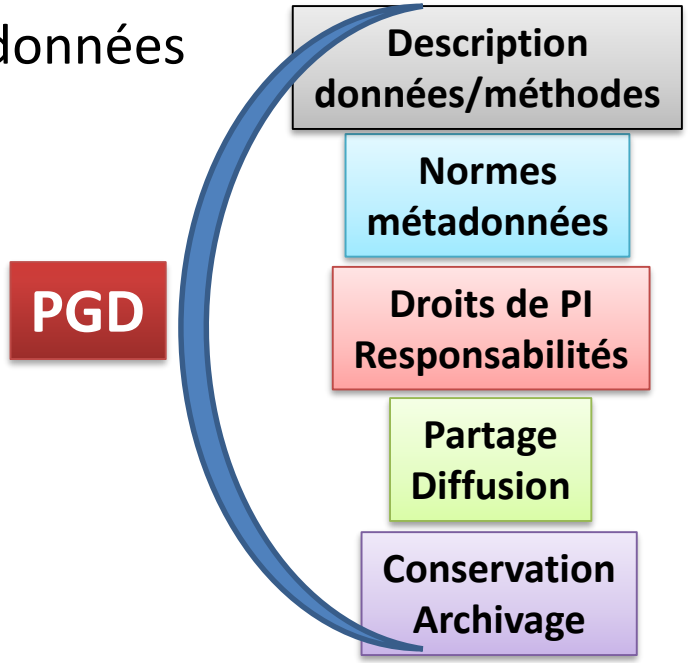
Faciliter la réutilisation des données implique la mise en œuvre de bonnes pratiques

- De **description et documentation**,

  en accord avec les pratiques de votre communauté scientifique:

  - protocoles, méthodes, unités de mesure
  - normes et standards de métadonnées
  - vocabulaires contrôlés,

- De **gestion** des fichiers
- De **conservation** des données
- De **traçabilité des droits**



2. Les enjeux du PGD

Le PGD concerne tous les producteurs de données

- Les scientifiques
  - surtout si départ en retraite / changement d'activités
- Les doctorants

→ pour laisser des données compréhensibles et réutilisables par d'autres

- Les projets de recherche et toutes collaborations
  - travail en équipe, projet en partenariat
  - multi-sites, multidisciplinaire, multi-traitements

→ pour harmoniser/standardiser les protocoles

2. Les enjeux du PGD

Le PGD comme outil d'animation de projet

Identifier les données à collecter/générer	Méthodes de collecte	Modes de description des données	Modes de stockage des données	Modalités de partage des données

- **1 PGD/projet** → vue d'ensemble + détails par jeux de données si besoin
- Préparer **dés le début du projet**, **avec tous les partenaires**
  - harmonisation des expérimentations entre partenaires

    i*mportant si disciplines différentes*

    lors de comparaison entre sites, contextes, traitements
  - partage des protocoles et des pratiques
  - standardisation des normes/métadonnées, unités de mesures,…
  - facilite échange de données, combinaison/agrégation/Base Données

2. Les enjeux du PGD

Le PGD comme outil d'animation de projet

Identifier les données à collecter/générer	Méthodes de collecte	Modes de description des données	Modes de stockage des données	Modalités de partage des données
• Décrit ce qui est collecté ou mesuré et comment; • comment les données sont nettoyées • comment elles sont traitées		Incite à respecter les standards pratiqués dans la discipline: vocabulaires, normes, métadonnées, unités		Oblige à s'interroger sur ses droits + sur le respect de règles éthiques à prendre en compte
Facilite la compréhension et l'interprétation des données		Facilite l'interopérabilité les traitements et la réutilisation	Evite la perte de données, Assure le respect de reglementations (RGPD)	

cirad

2. Les enjeux du PGD

Pose de bonnes questions dès le début du projet

- Comment décrire au mieux les données pour assurer leur compréhension ?
- Quelles données conserver ? Comment ?
- Quelles données partager ? Avec qui ?
- Avons-nous les droits ?
- Aspects éthiques à prendre en compte ?

  données sensibles, personnelles,
  issues de ressources bio du sud, ou de connaissances traditionnelles
  concernant des espèces végétales ou animales menacées
  ayant un impact sur homme/environnement
- Combien ça va couter ?

A quelles questions répond un PGD ?

- Comment la gestion des données est-elle financée, en particulier à long terme ?

Ressources

- En quoi consiste le projet ?
- Qui sont les partenaires ?
- Quelle est la politique de gestion des données ?
- Qui est responsable de la gestion des données ?

Responsabilités dans le projet

- Quelles données seront produites/utilisées au cours du projet ? (type, format, volume et accroissement…).
- Comment seront-elles produites ou transformées ?

Collecte des données

- Comment, où, par qui, seront stockées, sauvegardées et sécurisées les données ?

Sauvegarde des données

- Qui pourra accéder aux données ? Les données seront-elles partagées ? publiées ? Avec qui ?
- Comment ?
- Dans quel délai ?
- Sous quelle licence ?

Accès et partage des données

- Comment les données seront elles identifiées, décrites ?
- Quels standards de métadonnées utilisera t'on ?
- Comment seront générées les métadonnées ?

Documentation des données

- Quel plan pour l'archivage et la préservation à long terme ?

Archivage et préservation des données

- Des données sensibles seront-elles produites ou utilisées ?
- Comment sera assurée leur anonymisation ?

Ethique

- Qui sera propriétaire des données produites ?
- Des données externes seront-elles utilisées ?

Propriété intellectuelle

Les plans de gestion de données - S. Cocaud et D. L'Hostis, INRA. URFIST Paris - 11/07/18

4

Issu de : https://fr.slideshare.net/IST_IRD/grer-ses-donnes-avec-un-plan-de-gestion-de-donnes-pgddmp

2. Les enjeux du PGD

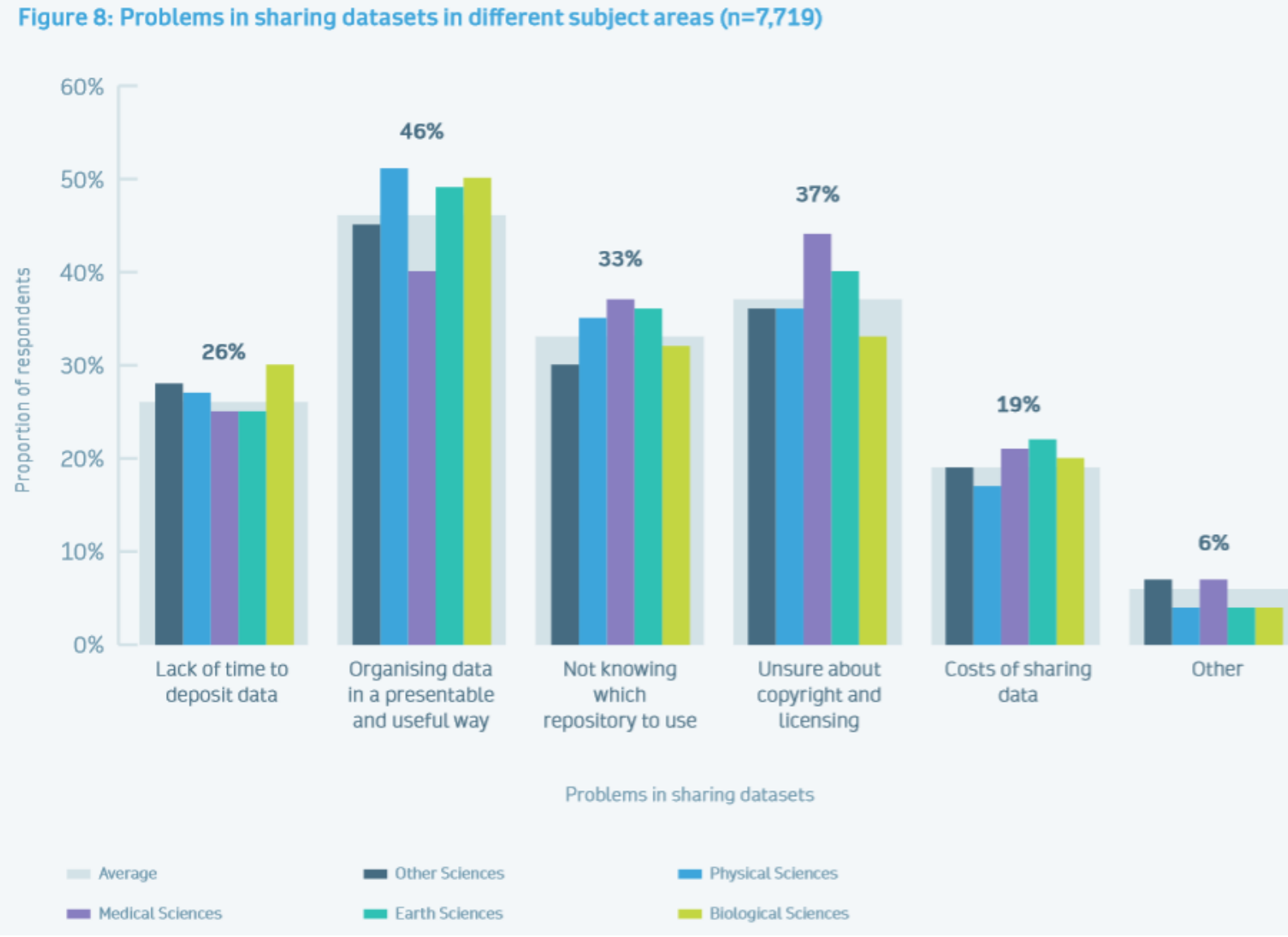
Etes-vous sûrs de pouvoir **comprendre et retrouver vos** derniers jeux de **données** ?

Par exemple :

- Ceux qui vous ont permis de publier un article de recherche
- Ceux que vous avez produits lors de votre dernier projet de recherche
- Ceux sur lesquels vous avez travaillé il y a 2 ans
- Ceux qui ont été produits par votre doctorant
- Etc.

→ **Questionnaire**

2. Les enjeux du PGD
→ Barrières au partage des données

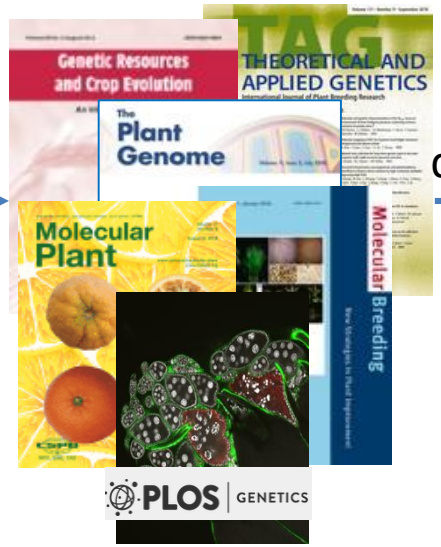


Springer Nature : *Practical Challenges for Researchers in Data Sharing* (mars 2018)
https://www.springernature.com/gp/open-research/open-data/practical-challenges-white-paper

cirad

2. Les enjeux du PGD
→ gain de temps pour publier

article



Dépôt données

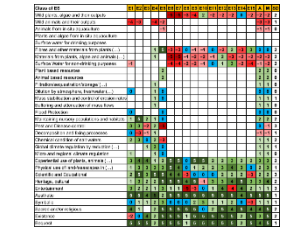
NCBI
GenBank
ENA
European Nucleotide Archive
EMBL-EBI
UniProt
European Variation Archive
ArrayExpress
GEO
Gene Expression Omnibus

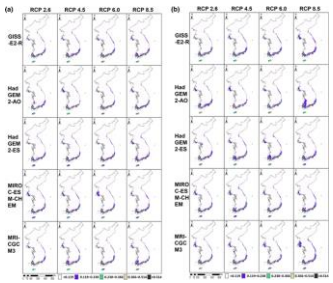
Standard de métadonnées

MIxS
MINSEQE
MIARE
MIMARKS
MIAME

Format

VCF, FASTA, FASTQ, MIQAS, CRAM, TABIX,…
CSV





article

Dépôt données

PANGAEA.
Data Publisher for Earth & Environmental Science
TreeBASE
DRYAD
figshare
EDI Data Porta
knb
LTER
Network Data Portal

Standard de métadonnées

EML
MIAPA
Darwin Core
ISO-8601
INSPIRE

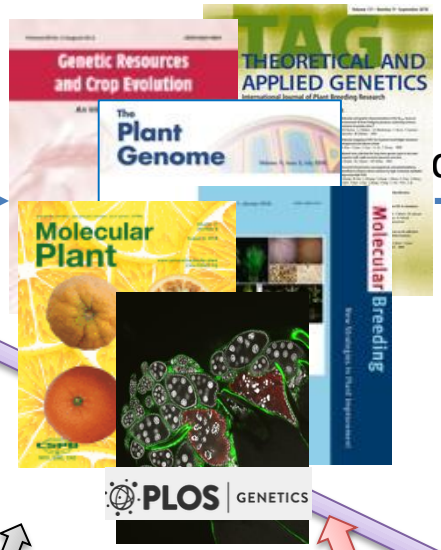
Format:

CSV, ASCII, NEXUS

cirad

2. Les enjeux du PGD
→ gain de temps pour publier

article



Dépôt données

NCBI
ENA
European Nucleotide Archive
GenBank
EMBL-EBI
UniProt
European Variation Archive
ArrayExpress
GEO
Gene Expression Omnibus

Standard de métadonnées

MIxS
MINSEQE
MIARE
MIMARKS
MIAME

Format

VCF, FASTA, FASTQ, MIQAS, CRAM, TABIX,… CSV

Description données et méthodes

Normes métadonnées

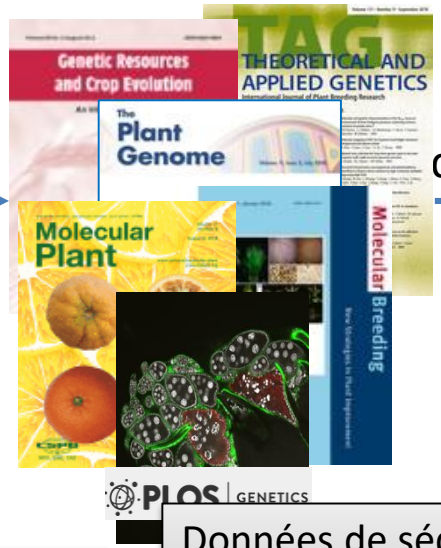
Droits de PI Responsabilités

Partage Diffusion

Conservation Archivage

2. Les enjeux du PGD
→ gain de temps pour publier

article



Dépôt données

NCBI
ENA
European Nucleotide Archive
GenBank
EMBL-EBI
UniProt
European Variation Archive
ArrayExpress
GEO
Gene Expression Omnibus

Standard de métadonnées

MIxS
MINSEQE
MIARE
MIMARKS
MIAME

Format

VCF, FASTA, FASTQ, MIQAS, CRAM, TABIX,…

Description des données	Données de séquençage de variétés de riz, cultivées dans X conditions, pour étudier…. Echantillons récoltés selon le protocole standard Y Séquences obtenues par Illumina HiSeq 2500,
Normes et métadonnées	Données documentées selon standard de métadonnées MIxS et stockées en format FASTA
Droits PI, Responsabilités	Toutes les données sont produites par le projet / par le Cirad. Tous les partenaires OK pour un libre accès aux données après publication Mr X, leader WP1, est responsable gestion données séquençage + Contact
Partage et diffusion	Les données seront déposées dans l'entrepôt ENA et seront accessibles sous licence CC-BY, après publication/à la fin du projet. Seules les données sur les variétés XX du riz seront accessibles, les autres étant obtenues sous contrat avec un partenaire privé….
Conservation et archivage	Données stockées sur ordi / mot de passe + sauvegarde hebdo sur disque externe + dépôt fichiers dans l'entrepôt institutionnel

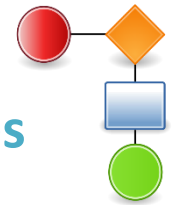
2. Les enjeux du PGD

https://www.youtube.com/watch?v=N2zK3sAtr-4&feature=youtu.be
4 min 40 sec

Pour le scientifique

- Mise en place de **bonnes pratiques** de gestion et documentation de ses données
- Réflexion sur les aspects **éthiques et juridiques**
- **Valorisation des données** : *datapaper*, stockage en entrepôt, mise en catalogue, etc.
- Si vous **partagez vos données** : + de visibilité, + de citations, + de notoriété et de nouvelles collaborations
- Si vous avez **accès à des données fiables** qui remplacent 2 ans d'expérimentation : gain de temps, gain financier et gain scientifique

Pour le collectif projet



- Offre une **vue globale** du projet et **flux de données**
- Favorise et facilite le **travail collaboratif**
- Outil de **support du management transversal**

Pour l'institution

- Participe à la mise en œuvre d'une **politique de partage** des données (**Open Science**)
- Constitution de **catalogues de données** réutilisables

Pour le scientifique

- Nécessite du temps (réflexion, rédaction, …)
  → **modèle documenté, outil d'aide à la rédaction existants**
- Crainte récurrente par rapport à l'Open Data
  → **Un PGD n'oblige pas à la diffusion des données**
  → **Conditions de réutilisation au choix de celui qui publie**
- Incompréhensions : « Comment décrire un jeu de données encore non produit ? »
  → **Peu de métadonnées servent à décrire un jeu de données**
- Désintérêt : perçu comme une tâche administrative
  → **« Outil de travail qui facilite la gestion et le partage des données tout au long du projet de recherche » (S. Pamerlon, GBIF)**

Pour le collectif projet

- Définir la granularité des jeux de données
  → **Spécifique à chaque projet !**
  → **Pistes de réflexion : potentiel de réutilisation des données**

- Susciter l'intérêt des membres du projet…
  → **Accompagnement : assistance, formation, organisation d'ateliers…**

- … tout au long du projet !
  → **Animation continue, mise en place d'outils collaboratifs**

3

REDIGER LE PGD

3. Rédiger le PGD

- **1 seul PGD par projet**
  - donne une vue d'ensemble du projet
  - décrit tous les jeux de données produits dans le projet
  - décrit le cycle de vie des jeux de données : de la collecte au partage
  - pendant et après le projet.
- Objectifs:
  - **compréhension, interopérabilité, réutilisation des données**
  - incite à utiliser des méthodes, protocoles, normes, métadonnées « standards » = reconnus dans vos disciplines
  - **compatibilité avec les « bonnes pratiques » de vos communautés**.
- **Première version** à délivrer au **mois 6** du projet (H2020)
  - pas d'obligation de répondre à toutes les questions posées
- Document **évolutif** : V1 (6 mois), V2 (18 mois), V3 (fin)

Structure du PGD - Séquences de la formation

Description données/méthodes

Normes métadonnées

Conservation Archivage

Droits de PI Responsabilités

Partage Diffusion

- **Décrire ses données**
  - Description des données (type, localisation, période, volume, …)
  - Méthodes de production
  - Documentation des données (normes, standards métadonnées)
- **Stocker et conserver ses données**
  - Modes de stockage, sécurisation et d'archivage
  - Gestion des fichiers, formats des données, outils de traitement
- **PI, réglementations, RGPD, licences de diffusion**
  - Droits de PI
  - Données perso, sensibles, issues de ressources génétiques
  - Réutilisation données existantes
- **Partage et diffusion des données**
  - Modes de partage pendant et après le projet
  - Potentiel de réutilisation et publics cibles
  - Entrepôts de données

3. Rédiger le PGD

PGD classique

- Description données/méthodes
- Normes métadonnées
- Droits de PI Responsabilités
- Partage Diffusion
- Conservation Archivage

Principes FAIR

Structuration des infos
pour favoriser
la découverte, l'accès
et la réutilisation des données
par l'homme (≠ acteurs)
et la machine (≠ systèmes).

DMP FAIR

- Data summary
- FAIR Data
- Making data Findable
- Making data openly Accessible
- Making data Interoperable
- Increase data Re-use
- Allocation of resources
- Data security
- Ethical aspects

Outils pour rédiger le PGD

Data cirad

https://intranet-data.cirad.fr/

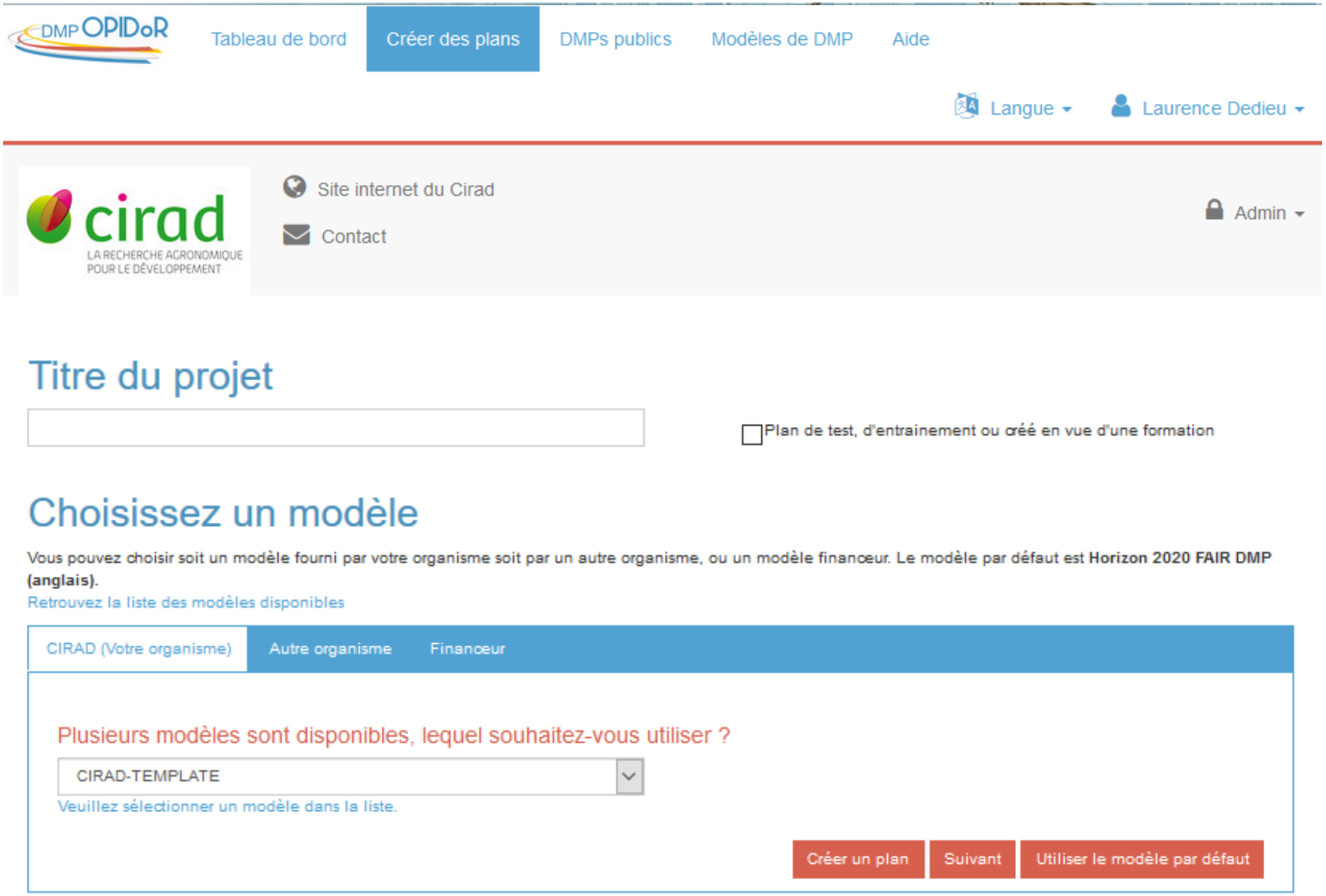
Outil d'aide à la rédaction du PGD

DMP OPIDoR

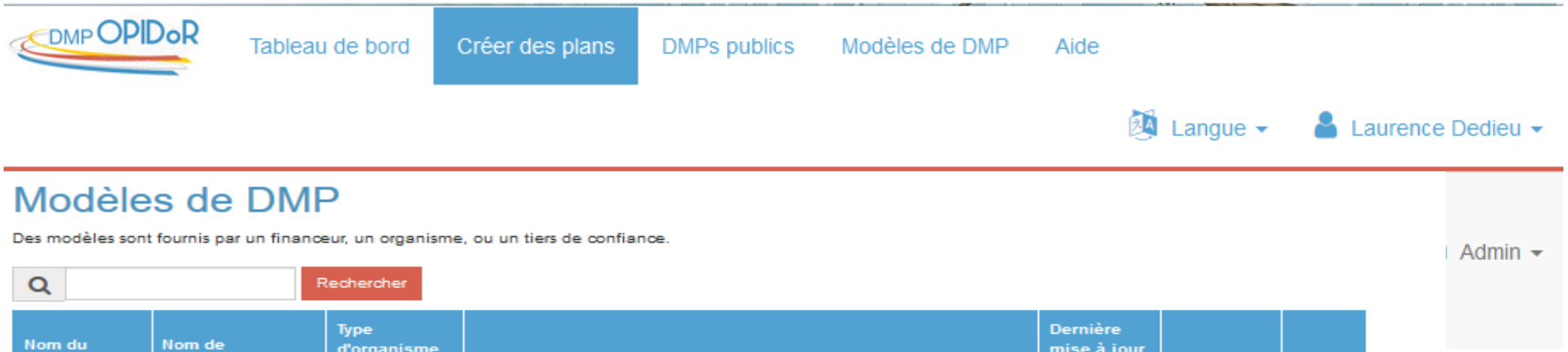
https://dmp.opidor.fr/

Les 2 modèles, PGD classique et FAIR, sont disponibles sur le site Data Cirad et sur DMP OPIDoR

Outils pour rédiger le PGD



Outils pour rédiger le PGD



Modèles de DMP

Des modèles sont fournis par un financeur, un organisme, ou un tiers de confiance.

Rechercher

Nom du modèle	Nom de l'organisme	Type d'organisme	Description	Dernière mise à jour	Télécharger	Actions
CIRAD-TEMPLATE	CIRAD	Etablissement	Modèle pour réaliser un Plan de Gestion des Données. Version spécifique CIRAD basée sur le modèle H2020	20/06/2018	DOCX PDF	
CIRAD-TEMPLATE-ENG	CIRAD	Etablissement	English version of the CIRAD model based on H2020 model for realizing a Data Management Plan.	05/11/2018	DOCX PDF	
Horizon 2020 DMP	Commission européenne	Financeur	In Horizon 2020 a limited pilot action on open access to research data will be implemented. Projects participating in the Open Research Data Pilot will be required to develop a Data Management Plan (DMP), in which they will specify what data will be open. Other projects are invited to submit a Data Management Plan if relevant for their planned research. The DMP is not a fixed document; it evolves and gains more precision and substance during the lifespan of the project. The first version of the DMP is expected to be delivered within the first 6 months of the project. More elaborated versions of the DMP can be delivered at later stages of the project. The DMP would need to be updated at least by the mid-term and final review to fine-tune it to the data generated and the uses identified by the consortium since not all data or potential uses are clear from the start. The templates provided for each phase are based on the annexes provided in Guidelines on Data Management in Horizon 2020	20/06/2018	DOCX PDF	
Horizon 2020 FAIR DMP (anglais)	Commission européenne	Financeur	The Commission is running a flexible pilot under Horizon 2020 called the Open Research Data Pilot (ORD pilot). Projects participating in the pilot must submit a first version of the DMP (as a deliverable) within the first 6 months of the project. The DMP needs to be updated over the course of the project whenever significant changes arise. The European commission provides a DMP template, the use of which is recommended but not mandatory. That template has been translated by the Inist-CNRS and is available in DMP OPIDoR Further details are provided in the Guidelines on FAIR Data Management in Horizon 2020 (v.3, 26 July 2016).	31/10/2018	DOCX PDF	

cirad

3. Rédiger le PGD

Critères d'évaluation des reviewers (1)

- L'information contenue dans le PGD est-elle appropriée au projet de recherche ?
- Le PGD vous semble t-il réalisable et adapté à la gestion des données du projet ?
- La pre-existence de données a t'elle été évaluée avant de décider d'en créer de nouvelles ?
- Le PGD couvre t-il bien toutes les données que le projet va produire ?

3. Rédiger le PGD

Critères d'évaluation des reviewers (2)

- Le PGD décrit-il suffisamment comment les données sont collectées, stockées (type de format) et documentées ?
- Les responsabilités ont-elles été nominativement attribuées ? Y compris pour les partenaires produisant des données ?
- Des données personnelles ou sensibles seront-elles collectées ? les mesures de sécurité appropriées ont-elles été prévues ?
- Le PGD est-il conçu pour maximiser le partage des données ?
- Tous les obstacles au partage ont-ils été pris en compte ?

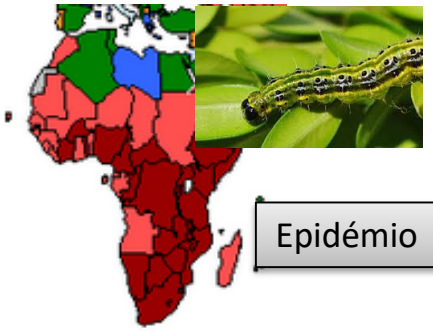
4

DÉCRIRE SES DONNÉES

4. Décrire ses données

OMICS

Biochimiques / sensorielles



Epidémio

Le PGD doit décrire tous les jeux de données produits dans un projet

Agro / physio / phéno

Climatiques

Sols / hydro

Enquêtes

4. Décrire ses données

Projet BREEDCAFS (*BREEDing Coffee for AgroForestry Systems)*

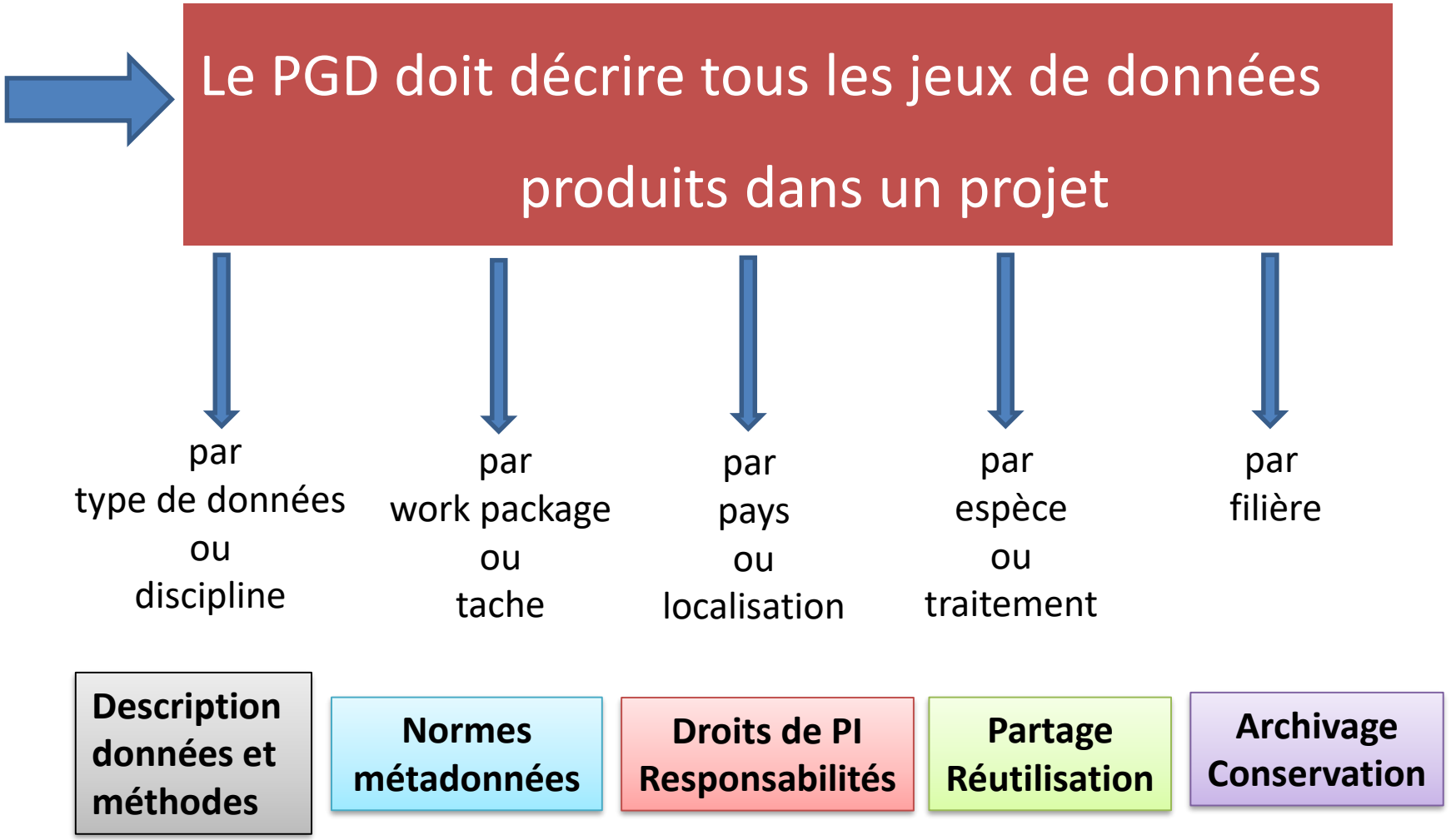
Diversité des types de jeux de données :

- Physiologiques (racines, feuilles, tiges, fruits des caféiers)
  - phénotypiques
  - morphométriques
- Génétiques et épigénétiques
- Transcriptomiques
- Biochimiques
- Performance agronomique
- Sensorielles
- Enquêtes auprès des fermiers

5 pays/40 sites
X hybrides + variétés

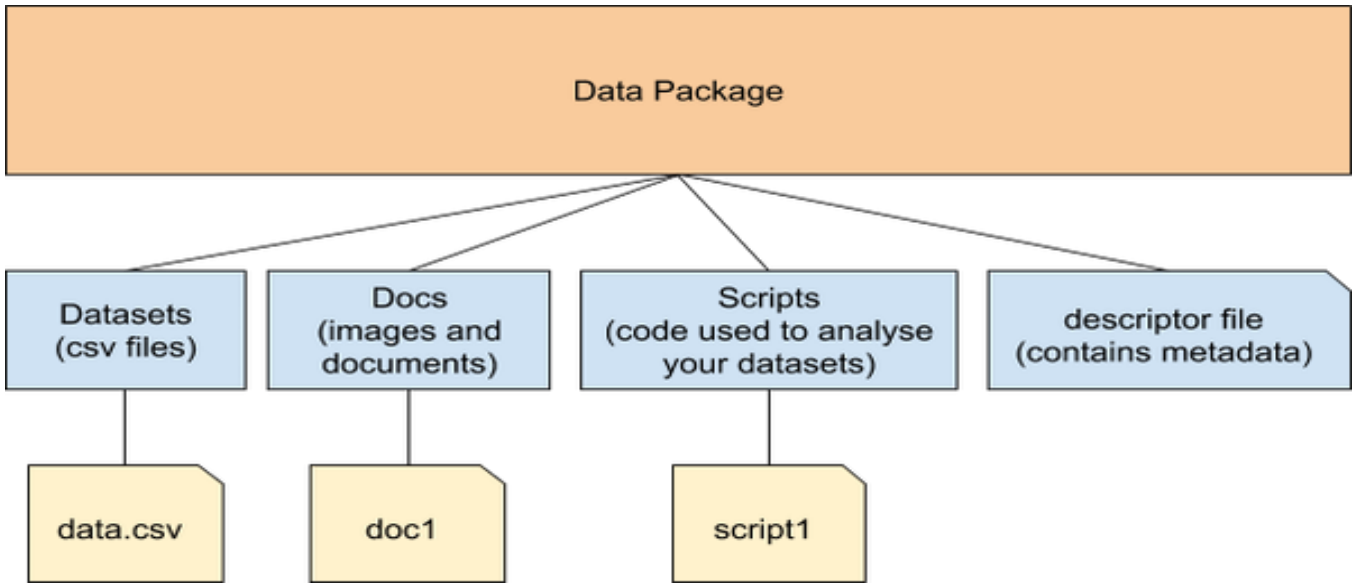
BREEDCAFS Database

4. Décrire ses données



4. Décrire ses données

Faciliter la réutilisation des données:
donc les rendre compréhensibles et interprétables



Standards de normes et métadonnées

Méthodes, protocoles, plan échantillonnage

Description des variables, Unités de mesure, abréviations

Equipement utilisé, méthode de calibration, contrôles

Questionnaires, guides d'enquêteurs, technique d'anonymisation

Explication du nommage de vos fichiers et des différentes versions

4. Décrire ses données

Données sur la composition chimique de feuilles de cacaoyers collectées en Côte d'Ivoire

	A	B	C	D	E	F	G	H	I	J
1	**ID**	**Lon**	**Lat**	**region**	**Produc**	**N**	**P**	**K**	**Ca**	**Mg**
2	Abe-1	-3.640017	6.712517	Abengourou	464	2,15	0,17	2,28	0,85	0,56
3	Abe-10	-3.695350	6.660200	Abengourou	763	2,21	0,13	2,94	1,66	0,56
4	Abe-2	-3.640050	6.710700	Abengourou	429	2	0,13	2,29	1,55	0,53
5	Abe-244	-3.248250	6.727933	Abengourou	600	2,21	0,14	2,6	1,11	0,48
6	Abe-245	-3.304450	6.653750	Abengourou	714	2	0,14	1,28	1,79	0,63
7	Abe-247	-3.300417	6.611333	Abengourou	625	1,89	0,15	2,69	1,53	0,72
8	Abe-268	-3.635183	6.568350	Abengourou	300	2,15	0,15	2,54	0,99	0,49
9	Abe-270	-3.659550	6.655417	Abengourou	650	1,78	0,17	3,03	0,67	0,56
10	Abe-271	-3.645900	6.673800	Abengourou	530	2,36	0,15	2,41	1,04	0,57
11	Abe-272	-3.701567	6.667383	Abengourou	355	2,05	0,19	2,2	1,16	0,71
12	Abe-3	-3.638017	6.706067	Abengourou	349	1,94	0,15	2,24	1,55	0,7
13	Abe-4	-3.645217	6.699017	Abengourou	256	2,1	0,15	2,72	1,32	0,57

Plantations de cacaoyers adultes
Echantillonnage en mars 2015

ID= Identifiant de chaque échantillon
Localisation géographique
Produc = yield of dry beans (kg/ha)
N, P, K, Ca, Mg : %

? Est-ce suffisant pour comprendre et réutiliser les données ?

4. Décrire ses données

- **Avec la documentation associée**
  - Méthodes, protocoles, plan échantillonnage → fichier « Lisez moi »
  - Dictionnaires des variables, Unités de mesure, abbréviations
  - Equipement utilisé, méthode de calibration, contrôles
  - Questionnaires, guides d'enquêteurs, technique d'anonymisation
  - Schéma de la base de données, fichiers de syntaxe
  - Liens vers les articles publiés à partir de ces données

- **Avec des métadonnées** (éléments descriptifs structurés)
  - ≠ **standards** de métadonnées
    - Génériques
    - disciplinaires

Richesse des métadonnées + documentation complète
→ assure la compréhension et réutilisation de vos données

Le standard de métadonnées génériques

Dublin Core : 15 métadonnées de base

DC Element	Notes
Title	Title of Data Collection
Creator	Authoring Entity of Data Collection
Subject	Keyword(s)
Description	Abstract
Publisher	Producer of Data Collection
Contributor	
Date	Production Date - Data Collection
Type	Kind of Data
Format	Type of File
Identifier	ID Number - Data Collection
Source	Sources - Used for Data Collection
Language	
Relation	Other Study Description Materials
Coverage	Time Period Covered
	Country
	Geographic Coverage
Rights	Copyright - Data Collection

Peu d'infos structurées décrivant les données

Habitat, climat ?
Mode de culture ?
Traitement ?
Plante ? Organe ?
Espèce ? Genre ?
Souche ?

+ les métadonnées sont enrichies
+ les données sont :
trouvables
compréhensibles
réutilisables

4. Décrire ses données

Connaissez-vous la **norme pour décrire une date** ou une période ?

Connaissez-vous des **standards de métadonnées** communément utilisés **dans votre domaine** ?

Normes et standards de métadonnées

- **Date et heure :** Norme ISO-8601

  pour lever l'ambiguïté quand les dates sont exprimées en chiffres
  - AAAA-MM-JJ
  - HH :MM :SS

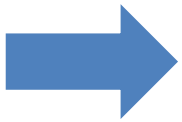
- **Géolocalisation**
  - Norme ISO 3166 (ex: BE pour Belgique, FR pour France…)
  - GeoNames: la base de données géographiques

  - Geospatial metadata standard, norme ISO-19115
  - Norme européenne INSPIRE

- **Propriétés chimiques des sols**
  - Normes ISO 13.080 - Qualité du sol. Pédologie
  - ISO 10694: Dosage carbone organique…; ISO 13878: teneur totale en azote….
  - ISO 28258: échange numérique des données relatives au sol.

Normes et standards de métadonnées

- **Phénotypage de plantes**
  - MIAPPE: Minimum Information About a Plant Phenotyping Experiment
- **Sciences Humaines et Sociales**
  - DDI : Data Documentation Initiative
- **Ecologie, Biodiversité**
  - EML : Ecological Metadata Language
  - Darwin Core
- **Génomique**
  - MIxS standards = Minimum Information about a *Genome Sequence, MARKer gene Sequence, Microarray*
    → déclinés pour 14 types d'environnement (Plant, soil, air, water, sediment,…)
  - MINSEQE = Minimum Information About a Next-generation Sequencing Experiment



- Utiliser un standard reconnu dans la discipline facilite l'interprétation de vos données pour un futur utilisateur
- Mentionner ce standard suffit pour le PGD

Le standard de métadonnées génériques

Dublin Core : 15 métadonnées de base

DC Element	Notes
Title	Title of Data Collection
Creator	Authoring Entity of Data Collection
Subject	Keyword(s)
Description	Abstract
Publisher	Producer of Data Collection
Contributor	
Date	Production Date - Data Collection
Type	Kind of Data
Format	Type of File
Identifier	ID Number - Data Collection
Source	Sources - Used for Data Collection
Language	
Relation	Other Study Description Materials
Coverage	Time Period Covered
	Country
	Geographic Coverage
Rights	Copyright - Data Collection

Peu d'infos structurées décrivant les données

Habitat, climat ?
Mode de culture ?
Traitement ?
Plante ? Organe ?
Espèce ? Genre ?
Souche ?

+ les métadonnées sont enrichies
+ les données sont :
trouvables
compréhensibles
réutilisables

MIAPPE: standard pour phénotypage plantes

Minimum Information About Plant Phenotyping Experiment

http://cropnet.pl/phenotypes/wp-content/uploads/2016/04/MIAPPE.pdf

- **Localisation**
  - Lieu géographique + lat / long / alt, habitat, …
- **Sujet**
  - Nom organisme, espèce, rang, génotype, âge, étape du cycle de vie, …
  - Semences: source, préparation, pretraitements, conservation
- **Environnement**
  - Culture: phytotrons, serre, jardin expé, champ,
  - Conditions: CO2, nutriments, eau, salinité, …
- **Traitements**
  - Chimiques, T°, maladies, fertilisants, hormones, radiations, pH,…
- **Sample collection, processing, management**
  - Organe/produit de la plante, traitement échantillon, T° conservation, …
- **Observed variables**
  - Variables phénotypiques et environnementales (trait, méthode, échelle)

MINSEQE: standard pour séquençage haut-débit

Minimum Information About a Next-generation Sequencing Experiment

https://www.ncbi.nlm.nih.gov/geo/info/seq.html#metadata

- **Description de l'échantillon**

  Nom de la source (ex : type de cellule) ; Organisme
  ≠ caractéristiques (ex: tissu, souche, genotype, phenotype, rRNA ratio, etc.)

- **Description des protocoles**

  d'extraction ; construction de la librairie ; technique de séquençage
  de croissance, traitement, conservation….

- **Etapes de traitement des données**

Gene Expression Omnibus

ArrayExpress

NCBI SRA

17	SAMPLES					
18	# This section lists and describes each of the biological Samples under investgation, as well as any protocols that are specific to individual Samples.					
19	# Additional "processed data file" or "raw file" columns may be included.					
20	Sample name	title	source name	organism	characteristics: tag	characteristics: tag
21	Sample 1					
22	Sample 2					
23	Sample 3					
24						
25	PROTOCOLS					
26	# Any of the protocols below which are applicable to only a subset of Samples should be included as additional columns of the SAMPLES section instead.					
27	growth protocol					
28	treatment protocol					
29	extract protocol					
30	library construction protocol					
31	library strategy					
32						
33	DATA PROCESSING PIPELINE					
34	# Data processing steps include base-calling, alignment, filtering, peak-calling, generation of normalized abundance measurements etc...					
35	# For each step provide a description, as well as software name, version, parameters, if applicable.					
36	# Include additional steps, as necessary.					
37	data processing step					
38	data processing step					
39	data processing step					
40	data processing step					
41	data processing step					
42	genome build					
43	processed data files format and content					
44						
45	# For each file listed in the "processed data file" columns of the SAMPLES section, provide additional information below.					
46	PROCESSED DATA FILES					
47	file name	file type	file checksum			
48						

INSPIRE: Interopérabilité des données géographiques

Description

*Intitulé de la ressource :

*Résumé de la ressource :

*Identificateur de ressource unique

*Catégorie thématique (1) :
(2) :
(3) :

*Thème INSPIRE :

*Rectangle de délimitation géographique :

Rectangle de l'emprise des données en degrés décimaux (par défaut, Fran

Région :	
Département :	
Commune (A-L) :	
Commune (L-Z) :	
Lat N / S	51,09
Long O / E	-5,79

*Référence temporelle

Date de la ressource (création): jj/mm/aaaa

Date de la ressource (publication): jj/mm/aaaa

Date de la ressource (dernière révision): jj/mm/aaaa

Etendue temporelle : (début) jj/mm/aaaa
(fin) jj/mm/aaaa

*Généalogie de la ressource

*Contraintes en matière d'accès et d'utilisation de la ressource

Limitations d'accès public

Restrictions d'accès public au sens d'INSPIRE	Valeurs autorisées mais insuffisant à établir la base légale des limitations d'accès public	contraintes de sécurité

Les conditions d'accès et d'utilisation décrivant les conditions applicables à l'accès et à l'utilisation des séries et des services de données géographiques, et, le cas échéant, les frais correspondants. Si aucune condition ne s'applique à l'accès à la ressource et à son utilisation, on utilisera la mention «aucune condition ne s'applique». Si les conditions sont inconnues, on utilisera la mention «conditions inconnues».

Conditions applicables à l'accès et à l'utilisation de la ressource :

Informations complémentaires sur la ressource

*Langue décrivant la ressource : français

Jeu de caractères de la ressource :

Type de représentation spatiale :

*Référentiel de coordonnées :

Version du format*

Encodage de la ressource :

Système de référence temporelle : grégorien

Cohérence topologique :

*Résolution spatiale

Résolution indiquée en echelle : 1/

OU Résolution indiquée en taille de pixels (mètres) :

Darwin Core : standard en biodiversité

- **Classification**, … , Vernacular name, …
  - “class”, “order", “family", “genus”, “subgenus”
- **Basis of Record**
  - Preserved Specimen, Fossil Specimen, Living Specimen
  - Human Observation, Machine Observation
- **Sex**
  - "female", "hermaphrodite", "male"
- **lifeStage**
  - "egg", "eft", "juvenile", "adult", "2 adults 4 juveniles“
- **establishmentMeans**
  - "cultivated", "invasive", "escaped from captivity", "wild", "native“
- **habitat**
  - "oak savanna", "pre-cordilleran steppe"

GBIF Global Biodiversity Information Facility

Freshwater Biodiversity Data Portal
The gateway to freshwater biodiversity data

A peer-reviewed open-access journal
Biodiversity Data Journal

Freshwater Metadata Journal

DwC : permet de rassembler des millions de données d’occurrences d’espèces dans le portail GBIF

EML: Ecological Metadata Language

Class II. Research origin descriptors

Site Description
- Site type
- Geography (location, size)
- Habitat
- Geology, Landform
- Climate

Experimental or sampling design
- Design characteristics
- Variables included
- Species sampled
- Data collection period, frequency

Research methods
- Field/Laboratory
- Instrumentation

SANParks
South African National Park Data Repository

The Knowledge Network *for* Biocomplexity
KNB

Class III. Data set status and accessibility

Status
- Latest update
- Latest archive date
- Metadata status
- Data verification

Accessibility
- Storage location and medium
- Contact person(s)
- Copyright restriction
- Costs

Class IV. Data structural descriptors

Data Set Files
- Data set Identity
- Size
- Format and storage mode

EML : permet l'interopérabilité des données en écologie
facilite la recherche, interprétation, sélection et exploitation

DDI (Data Documentation Initiative) : norme en SHS

Description des données

- Unité d'observation
- Couverture géographique
- Couverture temporelle (*AAAA-MM-JJ*)
- Univers

Méthodologie et traitement

- Dimension temporelle de l'enquête
- Collecteur des données
- Autres contributeurs
- Producteur
- Méthode d'échantillonnage
- Mode et Fréquence de collecte
- Opérations de contrôle
- Pondération
- Nettoyage des données

Dataverse cirad

ICPSR

beQuali

Journal of open health data

DDI : permet la réexploitation des données d'enquêtes grâce à la précision des données de contexte.

4. Décrire ses données

Guides de standards de métadonnées

FAIRsharing.org
standards, databases, policies

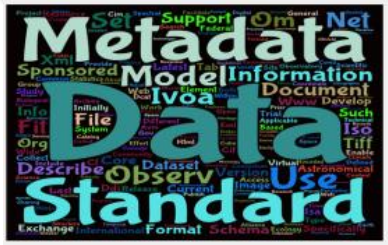
https://fairsharing.org/

elixir

https://www.elixir-europe.org/communities

http://www.dcc.ac.uk/resources/metadata-standards

Research Data Alliance



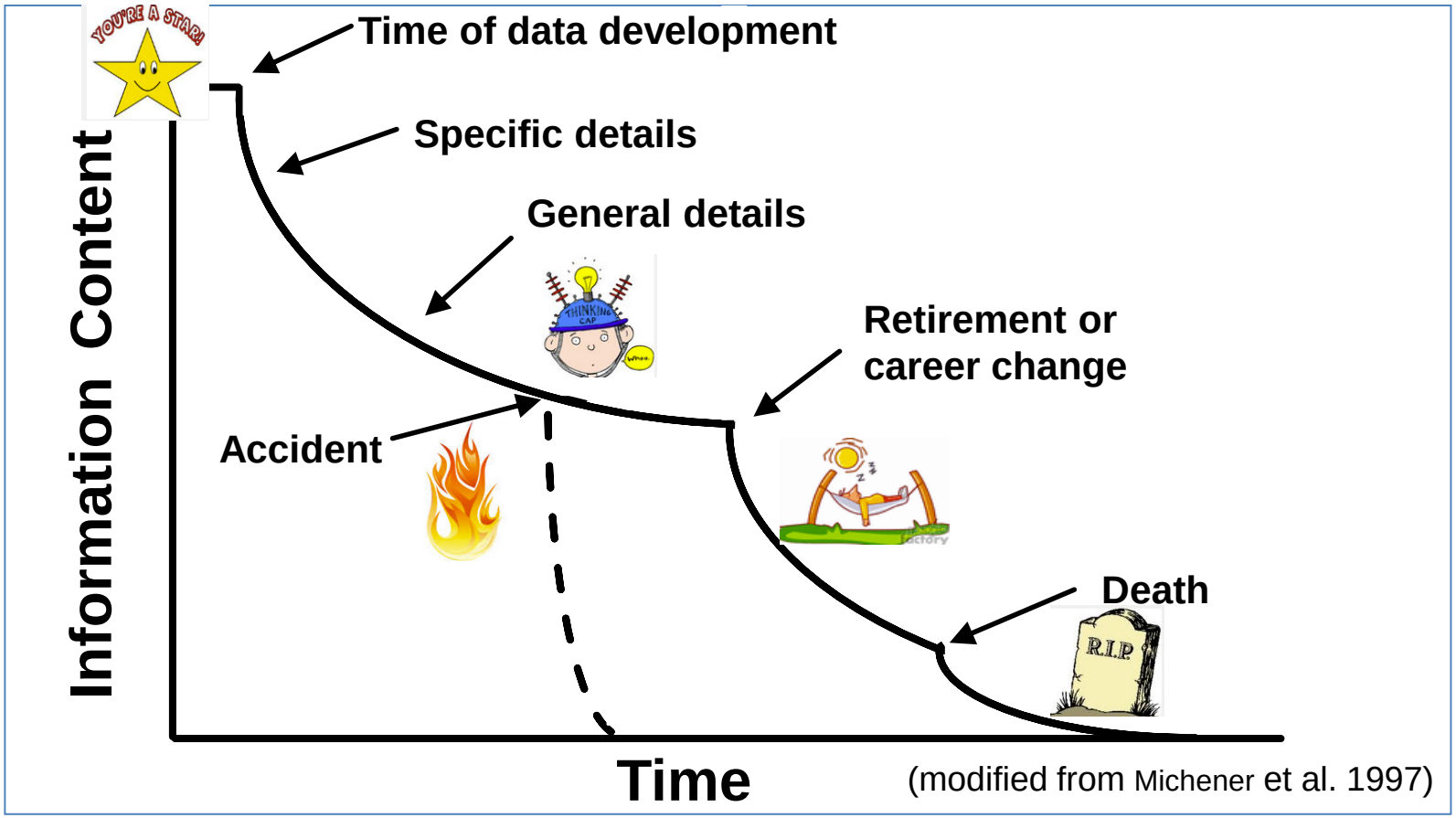
http://rd-alliance.github.io/metadata-directory/

ISO

https://www.iso.org/fr/home.html

- Tenir compte de l'entrepôt de données ciblé
  et de la revue (si publication prévue)
- Si aucun standard n'existe:
  expliquer comment vous créez les métadonnées ?

Metadata: Why are they important?



DataONE Education Module: Metadata. 2016.
http://www.dataone.org/sites/all/documents/L07_Metadata.pptx

4. Décrire ses données

- Objectif: **compréhension, interopérabilité, réutilisation** des données
  - utiliser des **méthodes et protocoles** classiques dans la discipline
  - décrire les données avec des **normes et métadonnées « standards »** = reconnus dans vos disciplines.

Méthodes standards + standards des métadonnées
+ Richesse des métadonnées
+ documentation complète

→ assure la compréhension
et la réutilisation de vos données

5

SÉCURISER, STOCKER ET ARCHIVER LES DONNÉES

5. Stocker, sécuriser et archiver ses données

Quelles sont vos pratiques en matière de : stockage, sauvegarde, sécurité et archivage de vos données ?

- D'après vous, quels risques pèsent sur les données pendant un projet?
- Où stockez les données ? Quelles sont les règles de sécurité ?
- Quelles données doivent être conservées ? Pendant combien de temps ?
- Quelles données doivent être supprimées ?

5. Sécuriser, stocker et archiver ses données
→ Evaluer les risques qui pèsent sur les données

Exemple de grille (1 = aucune gravité ; 5 = catastrophique)

	Gravité pour moi	Gravité pour les collègues	Gravité pour l'établissement	Gravité pour les participants	Gravité pour la société, l'environnement
A : Perte définitive	5	2	1	1	1
B : Indisponibilité temporaire	1				
C : Détérioration	2				
D : Accès non autorisé	1				
E : Modification non autorisée	2				
F : Lecture impossible	4				
G : Compréhension impossible	4				

Organiser, documenter et gérer ses données au quotidien – Mathieu Saby – mai 2019

5. Stocker, sécuriser et archiver ses données
→ Stockage des données

Fonction fondamentale : la **conservation des données**

Stockage

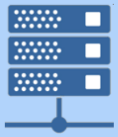
- désigne des méthodes et des technologies permettant de conserver des données
- concerne tout les types de supports de stockage de masse (DD, Clé USB…) ou support de stockage dématérialisé (cloud)
- répond a des problématiques d'usage collaboratif : dépôt, partage.

Critères de sélection pour choisir un support de stockage :

- la fréquence d'utilisation des données,
- les besoins en capacité de stockage (taille),
- La sécurité des données,
- la vitesse d'accès à la donnée
- La fiabilité et le cout du support

5. Stocker, sécuriser et archiver ses données

Comparatif de systèmes de stockage des données

Support de stockage	Sécurité	Accès	Coût	Remarque d'utilisation
 Ordinateur professionnel	★★☆☆ Sujet au piratage informatique, aux détériorations et pannes	★☆☆☆ Pas adapté au partage, nécessite l'utilisation d'un support externe ou d'Internet (mail, cloud…)	★★★★ Pas de coût supplémentaire ou coût peu important	- Pour un stockage temporaire - Nécessité de crypter les données confidentielles et sensibles
 Support externe	★☆☆☆ - Sujet au vol, à la perte du support - Durée de vie limitée (dégradation du matériel)	★★★☆ Facilement transportable, il permet de transférer les données vers un autre ordinateur	★★★★ Pas de coût supplémentaire ou coût peu important	- Pour un stockage temporaire - Nécessité de crypter ou de sécuriser physiquement les données confidentielles et sensibles
 Serveur institutionnel	★★★★ Stockage fiable, durable et sécurisé (contre le vol, le piratage, les incendies…)	★★☆☆ La connexion au serveur institutionnel ne facilite pas le travail avec des personnes extérieures	★★☆☆ Coût assez important mais pas forcément répercuté sur l'usager	- Pour un stockage plus pérenne - Adapté pour le stockage de données sensibles et des versions « stables » de vos données - Toutes les institutions ne proposent pas ce service
 Serveur Cloud	★★☆☆ On ne sait pas vraiment où sont stockées les données, ni ce qu'elles deviennent	★★★★ Permet un travail synchronisé avec toutes les personnes ayant été autorisées au partage	★★☆☆ Payant à partir d'une certaine limite de stockage	- Pour un partage avec des personnes externes à l'institution - Ne pas y mettre de données sensibles ou confidentielles - Pas de contrôle sur la procédure de sauvegarde des données

5. Stocker, sécuriser et archiver ses données
→ Sauvegarde Vs Archivage

Sauvegarde

Copie des données qui peut être utilisée pour restaurer les données originales dans le cas où ces dernières seraient endommagées ou perdues.

Après une sauvegarde, les données d'origine ne sont pas supprimées.

Le systèmes de sauvegarde servent à restaurer l'état des données à un instant T.

Archivage

L'archivage consiste à préserver les données :

- dans le respect de l'intégrité et l'authenticité,
- aussi longtemps que nécessaire (moyen et long termes),
- pour qu'elles soient en permanence accessibles, lisibles et compréhensibles.

La création d'une archive s'accompagne souvent de la suppression de l'original.

Les systèmes d'archivage servent à récupérer les données sur un intervalle de temps.

L'archivage peut être payant (coût éligible dans le projet de recherche).

5. Stocker, sécuriser et archiver ses données
→ Stockage sécurisé des données

Le stockage sécurisé est important pour assurer la continuité de l'exploitation des données sur du court terme.

Sécurité physique

Pour ne pas perdre ses données, la règle simple de la sauvegarde : 3-2-1

- Faire au moins 3 copies de ses données
- Stocker les copies sur au moins 2 supports différents
- Avoir au moins 1 sauvegarde hors site

- Zéro inquiétude pour la sauvegarde des données

Sécurité informatique

Pour se protéger des piratages et du vol de ses données, il est nécessaire :

- D'avoir un antivirus à jour pour minimiser les attaques informatiques

- De stocker ses données sur un ordinateur non connecté au réseau

- D'utiliser un mot de passe de niveau de protection élevé (longue chaine de caractère avec majuscule(s), chiffre(s), caractères spéciaux)

- De chiffrer ses données

Attention : perte définitive des données si mot de passe perdu

5. Stocker, sécuriser et archiver ses données
→ Recommandations pour le nommage des fichiers

Il est important d'avoir une stratégie de nommage des fichiers de données et de la respecter :

- Pour récupérer et identifier plus facilement les fichiers de données recherchés
- Éviter les problèmes lors de transfert et de partage
- Permettre leur conservation à moyen et long termes.

Quelques bonnes pratiques de nommage des fichiers :

- Choisir un nom court (- de 100 caractères) et significatif
- Pas d'espace, pas d'accent, pas de caractères spéciaux
- Utiliser des abréviations standards (ex : CR pour compte rendu)
- Indiquer le numéro de version du fichier
- Indiquer le nom de l'auteur
- Renseigner la date et/ou l'heure
- Penser à la numérotation (saisir des 0 initiaux pour les tris)

5. Stocker, sécuriser et archiver ses données
→ Recommandations sur le format des fichiers

Pour faciliter le partage des données, privilégiez les formats ouverts (bien documentés et utilisables sans demander d'autorisation)

Format ouvert (format libre)	Format fermé (propriétaire)
Spécifications techniques publiques	Spécifications techniques non publiques
Spécifications techniques sans restriction d'accès ni de mise en œuvre.	Des restrictions légales s'opposent à son utilisation (droit d'auteur, copyright, brevet) L'utilisation peut être payante
Format indépendant du logiciel utilisé qui assure l'interopérabilité des données	Format lisible qu'avec un logiciel spécifique

Ex de formats libres : DOCX, ODT, XLSX, ODS, XML …

Ex de formats propriétaires : PSD, JPG, WMA, WMV, RAR …

https://facile.cines.fr/ service de validation des formats

5. Sécuriser, stocker et archiver ses données
→ Quelles données conserver ?

Utilité

- Potentiel de réutilisation pour de futures recherches
- Uniques, non reproductibles (ex: photos), ou difficilement reproductibles (ex: expérience scientifique couteuse)
- Validation des résultats de recherche

Ex : Les observations de phénomènes météorologiques ou les données de référence sur les séquences de gènes

Obligation juridiques

Pour justifier d'une action ou d'une activité lorsqu'il y a contentieux (dans le cadre d'une enquête publique, d'un litige, d'une enquête policière, ou d'un rapport contesté en justice)

Ex : données prouvant le changement climatique

Intérêt historique

Pour témoigner de l'activité d'un organisme, d'une personne, d'une équipe

Ex : Les données épidémiologiques relatives à la pandémie de grippe espagnole en 1918 encore analysées 100 ans plus tard pour en tirer des leçons

5. Sécuriser, stocker et archiver ses données
→ Quelles sont les données à ne pas conserver?

Données à caractères personnelles

Une fois arrivé au terme du délai de conservation, les données personnelles doivent être supprimées ou anonymisées.

Données de simulations ou de calculs numériques

Toutes les données autres que le code source, les conditions initiales des simulations et des données de vérification.

Données de mauvaise qualité

- **Non documentées** : pas de métadonnées
- **Incomplètes** : ne pas disposer de toutes les donnée dont on a besoin
- **Inexactes** : fautes d'orthographe, informations manquantes, vides
- **Non conformes** : les données ne répondant pas aux normes réglementaires
- **Non actualisées** : une donnée non mise à jour devient obsolète et inutile.
- **Indisponibles :** utilisant des formats de fichiers propriétaires obsolètes, éparpillées entre plusieurs supports
- **Non sécurisées :** les données laissées sans contrôle peuvent être piratées

5. Sécuriser, stocker et archiver ses données

Le Cirad propose…

Dataverse comme lieu de stockage et d'archivage des données

https://dataverse.cirad.fr/

6

BONNES PRATIQUES JURIDIQUES

6. Bonnes pratiques juridiques

Cadre juridique de l'open data en France

Open Data : politique **d'ouverture des données :**

par l'Etat et les collectivités publiques ➡ fonds publics/ bien commun.

- Nombreux textes de référence (loi CADA, Loi Valter, Loi république Numérique, Code de la recherche, Code des relations entre le public et l'administration, Code de la Propriété Intellectuelle, directive Inspire…) ;
- Sur le plan Juridique l'OPEN DATA est une renonciation des droits sur les bases de données

Evolutions : **Loi République Numérique** – 7 octobre 2016 :

- De la demande de communication à un **système de diffusion spontanée ;**
- Principe de **mise à disposition et de réutilisation gratuite de** l'information publique :

  Disparition de l'exception des établissements de recherche ;
- Entre administrations, il ne peut pas y avoir de redevance pour se communiquer des données ;
- **Libre réutilisation des données/écrits scientifiques** publiées issues des activités de recherche financées au moins par moitié par l'argent public ;
- Autorisation du **Text et data mining.**

6. Bonnes pratiques juridiques

Communication et diffusion

Données

Obligatoirement Communicables

- Données dont la publication présente un intérêt :
  - Economique
  - Social
  - Sanitaire
  - Environnemental
- Base de données
- Données relatives à un risque impactant les populations
- Données géographiques
- Le répertoire

Objectif : publication progressive des principaux documents administratifs puis de tous les documents d'ici 2 ans (2019)

Communicables au choix (pas obligatoire, pas interdit)

Selon choix établissement (stratégie d'établissement)

Non Communicables

Données portant atteinte :

- Au secret de la défense nationale
- A la conduite de la politique extérieure de la France
- A la sureté de l'Etat
- Au secret protégé par la loi
- Au Secret industriel et commercial

Données protégées par un droit de propriété intellectuelle

Données personnelles

6. Bonnes pratiques juridiques

Réutilisation des Données

Données communiquées/ diffusées

Principe :
Libre réutilisation des données publiques (communication obligatoire)

Exception :
Réutilisation des données soumise à autorisation

MINI

- ✓ **Pas de nécessité de licence :**
- ✓ Obligation de respect er les 4 conditions suivantes lors de la réutilisation
- Non altération ;
- Non dénaturation ;
- Source ;
- Date de dernière mise à jour.

MAXI

- ✓ **Licence obligatoire** : fixée par décret/homologué e par l'Etat ;
- Redevances;
- Restrictions liées à intérêt général et proportionnées;
- Ne pas restreindre concurrence ;
- Utilisation commerciale ou non ;
- …….

Conditions :

- Si droit des tiers ;
- Si droit de propriété intellectuelle

Exception : le droit du producteur de BDD ne peut être opposé par l'administration.

- Si données produites dans mission SPIC, soumise à concurrence.

Contexte :

- Mise en œuvre de la politique/stratégie d'établissement
- Respect des règles éthiques et déontologiques

6. Bonnes pratiques juridiques

Open Data

- Dans quels cas dois-je diffuser mes données en open data ?
- Si la loi me l'impose : loi Cada, loi Valter, loi pour une république numérique (Dite Loi Axelle Lemaire)
- Si le bailleur me l'impose : condition prévue dans les règles de participation
- Si le consortium me l'impose : les parties peuvent décider collectivement de diffuser les données en open data à l'issue du projet.

6. Bonnes pratiques juridiques

Focus sur les données personnelles

- Cadre règlementaire : Règlement Général sur la protection des données personnelles (RGPD) qui entre en vigueur en France le 25 mai 2018

- Qu'est-ce qu'une donnée personnelle ?

  Toute information se rapportant à une personne physique identifiée ou identifiable ; est réputée être une «personne physique identifiable» une personne physique qui peut être identifiée, directement ou indirectement, notamment par référence à un identifiant, tel qu'un nom, un numéro d'identification, des données de localisation, un identifiant en ligne, ou à un ou plusieurs éléments spécifiques propres à son identité physique, physiologique, génétique, psychique, économique, culturelle ou sociale.

6. Bonnes pratiques juridiques

Introduction de la notion de finalité

- Le RGPD introduit la notion de **finalités d'usage des données personnelles.**

- Les données personnelles sont collectées pour des finalités déterminées, explicites et légitimes, et ne peuvent pas être traitées ultérieurement d'une manière incompatible avec ces finalités ; le traitement ultérieur à des fins archivistiques dans l'intérêt public, à des fins de recherche scientifique ou historique ou à des fins statistiques n'est pas considéré comme incompatible avec les finalités initiales (limitation des finalités).

6. Bonnes pratiques juridiques

Focus RGPD

OPEN DATA =
LIBRE REUTILISATION
DES DONNEES

DONNEES
PERSONNELLES =
PRINCIPE DE
FINALITE DE
TRAITEMENT

ANONYMISATION
=
OUVERTURE
POSSIBLE

6. Bonnes pratiques juridiques

Focus sur les données sensibles

Qu'est-ce qu'une donnée sensible ?

Les données sensibles sont celles qui font apparaître, directement ou indirectement, les **origines raciales ou ethniques**, les **opinions politiques**, **philosophiques ou religieuses** ou **l'appartenance syndicale** des personnes, ou sont relatives à la **santé** ou à la **vie sexuelle** de celles-ci.

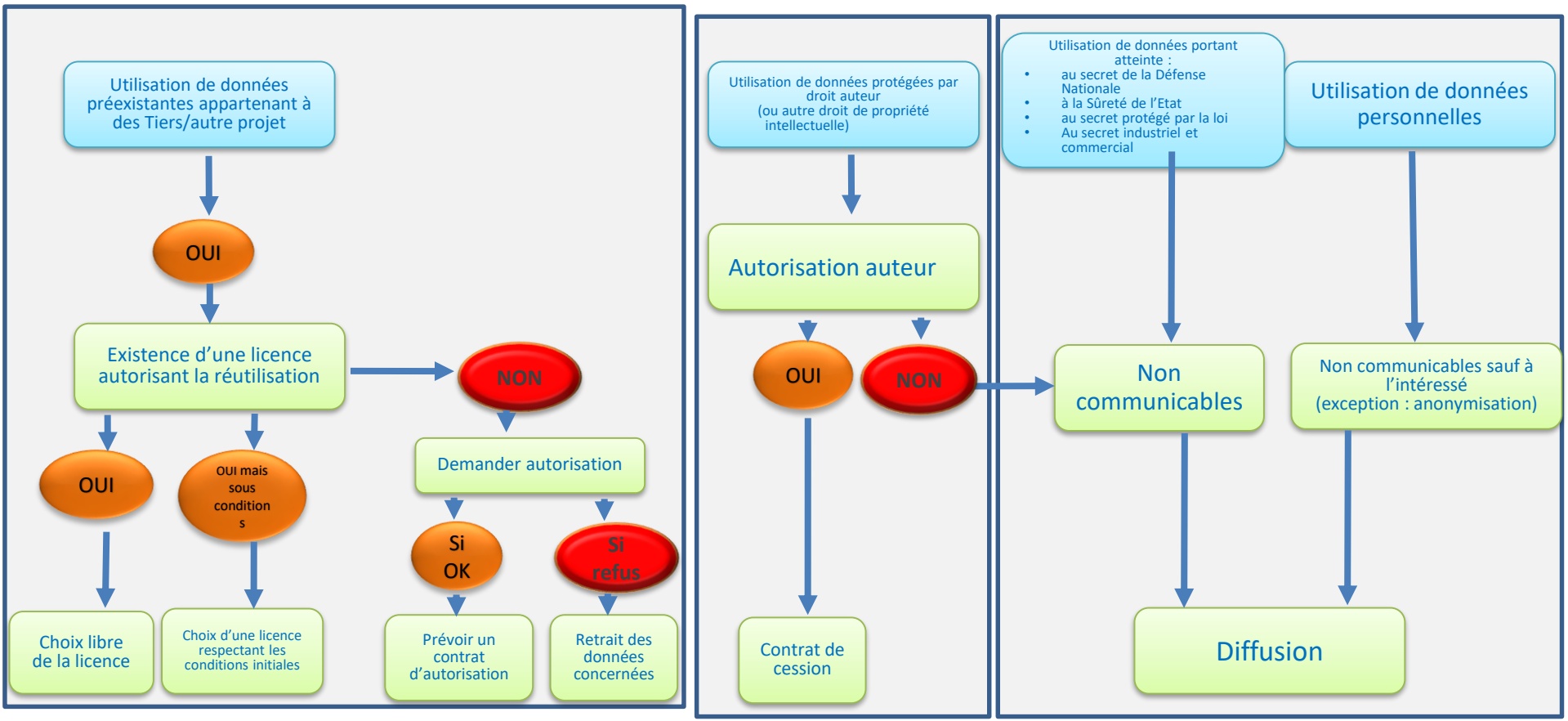
Par principe la collecte et le traitement de ce genre de données sont interdits.

Cependant dans la mesure où la finalité du traitement l'exige, ne sont pas soumis à cette interdiction :

- Les traitements pour lesquels la personne concernée a donné son consentement exprès
- Les traitements justifiés par un intérêt public après autorisation de la CNIL ou décret en Conseil d'État
- La collecte et le traitement de ces données doivent dans ces hypothèses être justifiées au cas par cas au regard des objectifs poursuivis

6. Bonnes pratiques juridiques

Données générées par le projet



6. Bonnes pratiques juridiques

http://www.bibliotheque-numerique.fr/DonneesDiffusables.php

Logigramme issu des travaux réalisés dans le cadre de la rédaction du guide sur l'Ouverture des données de la recherche, Guide d'analyse du cadre juridique en France

6. Bonnes pratiques juridiques

La licence – Outil de diffusion

- La licence est un contrat qui permet de déterminer les conditions de diffusion et de réutilisation des jeux de données/bases de données
- La licence peut permettre un simple droit d'accès pour consultation ou autoriser l'extraction des données
- La licence peut être ouverte (ex : licence Creative commons, ODBL…) ou fermée (licence propriétaire)
- Lors de la diffusion de la base de donnée, je m'interroge sur la licence sous laquelle je diffuse

6. Bonnes pratiques juridiques

La licence Creative Commons

Licence CC	Bouton	Explications	$ Exploitation permise?	Remix permis?
Attribution		• On doit citer **QUI** est l'auteur de l'oeuvre originale. • L'utilisation commerciale de l'oeuvre est permise. • Vous pouvez remixer l'oeuvre. • Le partage de l'oeuvre est permis.	☺	☺
Share Alike Partage à l'identique		• On doit citer **QUI** est l'auteur de l'oeuvre originale. • L'utilisation commerciale de l'oeuvre est permise. • Vous pouvez remixer l'oeuvre. • Cette licence doit toujours être utilisée sur toutes vos versions dérivées de l'oeuvre originale. • Le partage est permis.	☺	☺
No Derivative Modification non permise		• On doit citer **QUI** est l'auteur de l'oeuvre originale. • L'utilisation commerciale de l'oeuvre est permise. • Vous **NE** pouvez **PAS** remixer l'oeuvre. • Le partage est permis.	☺	☹
Non-Commercial Usage commercial non permis		• On doit citer **QUI** e st l'auteur de l'oeuvre originale. • L'utilisation commerciale de l'oeuvre **n'est PAS** permise. • Vous pouvez remixer l'oeuvre. • Le partage est permis.	☹	☺
Non-Commercial Usage commercial non permis + **Share Alike** Partage à l'identique		• On doit citer **QUI** est l'auteur de l'oeuvre originale. • L'utilisation commerciale de l'oeuvre **n'est PAS** permise. • Vous pouvez remixer l'oeuvre. • Cette licence doit toujours être utilisée sur toutes vos versions dérivées de l'oeuvre originale. • Le partage est permis.	☹	☺
Non-Commercial Usage commercial non permis + **No Derivative** Modification non permise		• On doit citer **QUI** est l'auteur de l'oeuvre originale. • L'utilisation commerciale de l'oeuvre **n'est PAS** permise. • Vous **NE** pouvez **PAS** remixer l'oeuvre. • Le partage est permis.	☹	☹

6. Bonnes pratiques juridiques

Outils d'aide au choix de licence

- http://licentia.inria.fr/licenseservice
- https://ufal.github.io/public-license-selector/
- https://eudat.eu/services/userdoc/license-selector

6. Bonnes pratiques juridiques

Exemples de licences ouvertes

- **Creative Commons**

- **ODBL** : Open Database License (ODbL) est un contrat favorisant la libre circulation des données.
  - Licence autorisant l'exploitation commerciale
  - Partage à l'identique des conditions initiales

- **Etalab** (FR) : facilite et encourage la réutilisation des données publiques mises à disposition gratuitement
  - Une licence ouverte, libre et gratuite ;
  - Une licence qui promeut la réutilisation la plus large en autorisant la reproduction, la redistribution, l'adaptation et l'exploitation commerciale des données ;
  - Une licence qui s'inscrit dans un contexte international en étant compatible avec les standards des licences Open Data développées à l'étranger (Open Government Licence, ODC-BY, CC-BY 2.0…).

7

PARTAGE ET VALORISATION DES DONNÉES

Questions posées dans le PGD

Pendant et après le projet

- Quels jeux de données ?
- Quand ?
- Où ? → quel entrepôt de données ?
- Comment ? → quelles modalités de partage

- Publics cibles ?
- Potentiel de réutilisation ?

Questions posées dans le PGD

Pendant le projet

- Avec qui ?
  - Tous les partenaires du projet
  - D'autres personnes / partenaires hors projet
- Où ?
  - Alfresco (documents)
  - Dataverse Cirad (données)
- Comment ? Modalités de partage et d'accès en interne
  - Les jeux de données seront déposés sur le Dataverse du Cirad et partagés avec tous les partenaires du projet.
  - Présence de partenaires privés dans le consortium → détails dans l'accord de consortium

7. Partage et valorisation des données

Partage fortement recommandé
(*institution, bailleurs, revues, …*)
parfois obligatoire (projets B&M Gates)

- Choix des jeux de données à partager

As open as possible, as closed as necessary

protection des données sensibles, personnelles, de santé, partenariats privés,… → à justifier dans le PGD

- Choix des modalités d'accès à vos données

licence ouverte/accès restreint/enregistrement
période d'embargo
limiter réutilisation (non commerciale, selon projet)

7. Partage et valorisation des données

Quels jeux de données partager ?

- **Données qui feront l'objet d'articles de recherche**
- Données ayant un potentiel de réutilisation
  - Nouvelles analyses, nouvelles questions de recherche
  - Méta-analyses, nourrir des modèles
  - Changement d'échelle : spatiale, temporelle, analyses stat.
  - Développement commercial, de services…
  - etc.
- Données utiles (jeux de données contrôles, lots témoins)
- Données présentant un intérêt pour certains publics

Identifier ses futurs utilisateurs

Quels publics seraient intéressés ?

- Communautés scientifiques
- Enseignants
- Décideurs
- Secteur privé, créateurs de start-up
- Grand public
- ONGs, associations internationales influentes
- Journalistes

7. Partage et valorisation des données

Cas justifiant une diffusion restreinte à expliciter

- Données issues de partenariats (ex: avec des acteurs privés)
- Données sensibles :

  concernant des espèces protégées ou envahissantes,
  données cliniques, issues d'expérimentations animales
  données économiques, personnelles
  issues de ressources biologiques du Sud (réglementation APA)

- Données stratégiques que vous souhaitez exploiter :

  identification de marqueurs génétiques, d'aromes
  création d'une appli, d'une base de données originale

- Jeux de données contenant des données préexistantes (produits par d'autres, sous licences non ouvertes, …)

Potentiel de réutilisation des données

Estimer la valeur de ses données

- Données rares ou uniques
  - Expérimentation impossible à répéter
  - groupes difficilement accessibles
  - phénomènes rares
- Données à forte valeur scientifique
  - données de référence
  - reproduction difficile ou couteuse
  - Ayant un grand intérêt pour certains publics (ex: société civile, pays du Sud)
- Données ayant une valeur économique
  - perspectives d'application, développement commercial
- Données ayant une valeur environnementale

7. Partage et valorisation des données

Quand partager ses données ?

- Vous avez exploité vos données
- Vous avez publié vos résultats de recherche
- Vous avez mis en forme vos données et métadonnées
- Vos données ont été anonymisées
- Vos partenaires sont d'accord
- Apres une période d'embargo
- Apres une période d'accès restreint ou sur demande

Où partager vos données ?

Dans un entrepôt de données
→ optimise les possibilités de réutilisation

- Entrepôt adapté à vos données
  - thématique, multidisciplinaire,
  - institutionnel, Européen
  - reconnu (notoriété) dans votre discipline
- Correspondant aux demandes : bailleur, institution, partenaires, revue où vous publiez
- Correspondant aux publics scientifiques visés
- Acceptant les modalités de diffusion que vous souhaitez
  - ⚠ qqs entrepôts en CC0
- Qui délivre un identifiant numérique pérenne et unique

Quels entrepôts de données ?

Connaissez-vous des **entrepôts de données dans votre domaine** ?

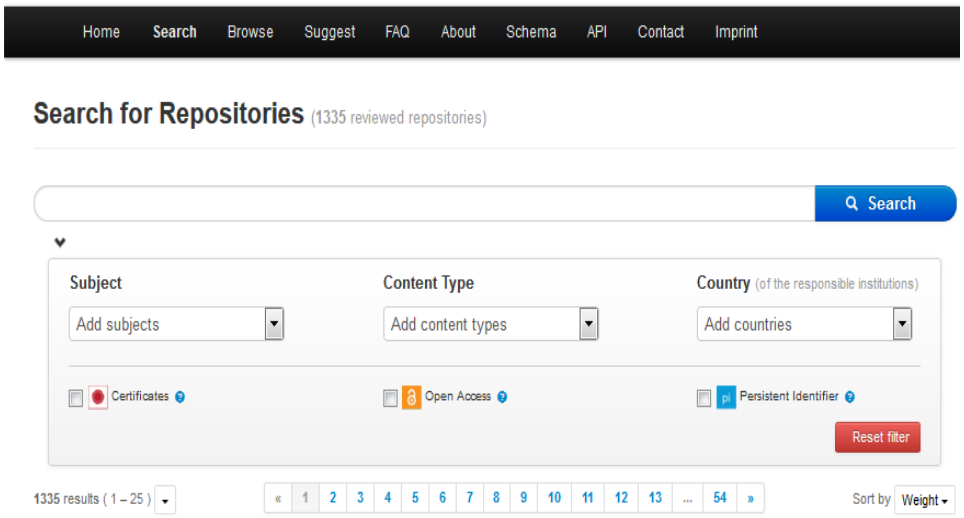
pour trouver des données fiables ?

pour déposer des données ?

Les entrepôts de données

re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES

→ 2000 entrepôts



Liste d'entrepôts de données (*Data repositories*)

dans les thèmes du Cirad

Laurence Dedieu, janvier 2018

Table des matières

Les entrepôts de données

- **Institutionnels** Dataverse
- **Europe** Zenodo, B2Share
- **Généralistes** Figshare, Dryad
- **Editeurs** Oxford Univ Press (GigaDB) ; Ubiquity Press (Dataverse)
- **Thématiques**
  - GBIF (Global Biodiversity Information Facility)
  - KNB (Knowledge Network for biocomplexity), EDI (Environmental Data Initiative)
  - Pangaea, SEANOE
  - Movebank, WormBase, ViPR, MycoBank, ComBase, FLOW
  - GenBank, Barcode of Life Data Systems, UniProt, Intact
  - TropGeneDB, Gramene, Plant Metabolic Network
  - Dataverse, ICPSR, DataFirst, Quetelet, beQuali

Optimiser la diffusion de ses données

Dupliquer

- Dépôt des données dans l'entrepôt du Cirad :

- Duplication de quelques métadonnées dans un entrepôt thématique partagé par votre communauté scientifique, avec le lien vers les données sur le Dataverse

Valoriser vos données

Faites savoir que vos données sont de qualité
qu'elles ont un potentiel de réutilisation
qu'elles sont disponibles

- Publier un **datapaper**
- Publier un **article de recherche**
- Rédiger une brève pour un **magazine** spécialisé
- Contribuer à un **blog**, ….

Journal of open health data

A peer-reviewed open-access journal
Biodiversity Data Journal
Making your data count! ISSN 1314-2828 (online)

Open Data Journal for Agricultural Research

cirad

7. Partage et valorisation des données
→ Le PGD peut être publié

Pensoft

RIO Research Ideas and Outcomes — The Open Science Journal

https://riojournal.com/

Data Management Plan: Empowering Indigenous Peoples and Knowledge Systems Related to Climate Change and Intellectual Property Rights
Cath Traynor
Data Management Plan | doi: 10.3897/rio.3.e15111
25-07-2017 | Unique: 296 | Total: 476 | Reprint: € 2,30 | See collection | HTML XML PDF

Data Management Plan: Opening access to economic data to prevent tobacco related diseases in Africa
Lynn Woolfrey
Data Management Plan | doi: 10.3897/rio.3.e14837
24-07-2017 | Unique: 242 | Total: 392 | Reprint: € 2,60 | See collection | HTML XML PDF

Data Management Plan: HarassMap
Reem Wael
Data Management Plan | doi: 10.3897/rio.3.e15133
18-07-2017 | Unique: 239 | Total: 374 | Reprint: € 2,60 | See collection | HTML XML PDF

Data Management Plan: IDRC Data Sharing Pilot Project
Cameron Neylon
Data Management Plan | doi: 10.3897/rio.3.e14672
27-06-2017 | Unique: 320 | Total: 537 | Reprint: € 2,60 | See collection | HTML XML PDF

Data Management Plan: Brazil's Virtual Herbarium
Dora Ann Lange Canhos
Data Management Plan | doi: 10.3897/rio.3.e14675
27-06-2017 | Unique: 360 | Total: 590 | Reprint: € 2,60 | See collection | HTML XML PDF

CPSE Labs

Deliverable 1.1

Data Management Plan

DISSEMINATION LEVEL		
PU	Public	X
CO	Confidential, only for members of the consortium (including the Commission Services)	

cirad

CONCLUSION

Le Plan de Gestion des Données

- Décrit **tous les jeux de données** produits dans un projet
  - Y compris ceux n'ayant pas vocation à être partagés
  - et leur gestion pendant et après le projet
- Incite à utiliser des **méthodes/protocoles, normes/métadonnées reconnus dans la discipline**
- Document **évolutif** : V1 (6 mois), V2 (18 mois), V3 (fin)
  - La V1 peut ne pas traiter toutes les questions
- À préparer **le plus tôt possible**
- **Outil d'animation** dans un collectif,
  - doit être partagé entre partenaires

Pour vous accompagner

Data cirad

https://intranet-data.cirad.fr/

Dataverse cirad

re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES

FAIRsharing.org
standards, databases, policies

Dist
De la recherche d'information à la publication
Faire une bibliographie
Rédiger et publier

https://intranet-dist.cirad.fr/

DCC because good research needs good data

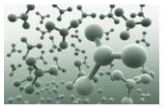
Home | Digital curation | About us | News | Events | Resources

Search by Discipline

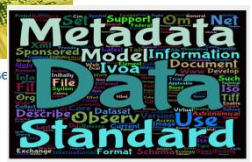

Biology


Earth Science


General Rese


Physical Science


Social Science & Humanities



CoopIST Coopérer en information scientifique et technique

https://coop-ist.cirad.fr/

cirad

FIN